PAPER

Evaluating Written Corrective Feedback:

A Comparative Study of Human-Written and Human-in-the-Loop 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

Written corrective feedback is central to developing a student's essay writing skills, but feedback requires extensive time and effort on the part of educators (Sia & Cheung, 2017). To optimize the efficiency and quality of feedback creation, Paper's Review Center has introduced a generative AI tool for use by Review Center tutors. This tool generates suggested written corrective feedback that tutors can accept, edit, or reject. Tutors use this tool in conjunction with their personally composed feedback comments. Thus, this tool pioneers a Human-in-the-Loop (HITL) integration for written corrective feedback. In the present report, we examine the optimization of *quality*. We demonstrate that the HITL approach in English language essay review significantly enhances the review quality compared to feedback given solely by human tutors.

We leverage LLM-based binary text classification to automatically evaluate whether or not in-line feedback comments are *encouraging*, *inquiry-based*, and *specific*. Quantitatively, we find that on average, the percentage of *encouraging*, *inquiry-based*, and *specific* comments per essay are significantly higher (9.9%, 12.6%, and 5.4% more comments, respectively) when tutors have access to a generative AI tool compared to when they do not.

Building off of these insightful quantitative results, we also conduct a granular qualitative analysis to bring to light general patterns, strengths, and limitations of AI and human written corrective feedback. We acknowledge that while AI-generated suggestions contain high proportions of *encouraging*, *inquiry-based*, and *specific* comments, human oversight is crucial to ensure AI capabilities are applied responsibly. Responsible use of AI positions HITL and human-centered AI systems as an ethical framework for AI usage in various domains, including educational technology (Chen et al. 2023; Klimova et al. 2023; Renz & Vladova 2021; Viola et al. 2023). To integrate AI in EdTech, a HITL approach provides tutoring professionals with tools that strengthen their abilities without ignoring their agency and expertise.

Data

Paper stores all written corrective feedback comments and AI suggestions historically used for their Review Center service. From this collection of data, we sampled 200 essays (submissions) reviewed by tutors in English *without* access to the generative AI tool. We also sampled 200 submissions reviewed by tutors in English *with* access to the generative AI tool.[1] In both datasets, most submissions are written by high school aged students. The full grade-level range includes grades 4 and beyond. All submissions are written in English.

Datasets	
Human Only Dataset (Submissions Reviewed without AI Access)	**HITL Dataset[2]** (Submissions Reviewed with AI Access)
● 1,791 comments ● On average, 9 comments per submission ● Most often, 5 comments per submission ● At least 1 comment per submission ● At most 31 comments per submission	● 2,387 comments ● On average, 12 comments per submission ● Most often, 8 comments per submission ● At least 4 comments per submission ● At most 40 comments per submission

The HITL Dataset has 4 comment types: Human Only, Accepted, Edited, and Rejected. 41.3% of HITL comments were solely produced by humans.

HITL Comment Types
Human Only: Comments produced without any AI suggestion (41.3%)
Accepted: AI suggested comments directly accepted by a human (31.9%)
Rejected: AI suggested comments rejected by a human (17.6%)
Edited: AI suggested comments edited by a human (9.2%)

[1] All AI-generated suggestions are not the output of one single prompt. In the present report, the term “AI” refers to a sample of aggregated outputs of our generative AI tool over the course of September 2023. During this time period, four unique prompts were used, with 77.5% of in-line comment feedback generated by the same prompt.
[2] Includes rejected suggestions.

Method

Experimental Groups

From our two datasets, we aggregate 9 groupings of comments for experimental comparison. A summary of these feedback slices are presented in the table below.

Experimental Groups		
Feedback Slice	**Description**	**Number of Comments**
Human Only Comments (No AI Access)	Comments by humans without access to AI tool	1,791
Human Only Comments (Had AI Access)	Comments by humans with access to AI tool	985
Rejected AI Comments	Rejected AI comment suggestions	420
Pre-Edited AI Comments	AI comment suggestions before editing	220
Edited AI Comments	AI comment suggestions after editing	220
Accepted AI Comments	Accepted AI comment suggestions	762
All Pure AI Comments	All comment suggestions by AI without human oversight (Accepted, Pre-Edited, Rejected)	1,402
Final AI Interacted Comments (Accepted, Edited)	Final AI interacted comment suggestions (Accepted, Edited) sent to students from tutors with access to AI tool	982
Final HITL Comments (Human, Accepted, Edited)	Full set of final comments sent to students from tutors with access to AI tool	1,967

Table 1. Table defining and quantifying the level of human-in-the-loop input within feedback comment types (feedback slices)

From the Human Only Dataset, we group all 1,791 comments as ***Human Only Comments (No AI Access)***. From the HITL Dataset, we group the 985 human only comments as ***Human Only Comments (Had AI Access)***. Comparing these two groups allows for examination of differences between the human tutors in either dataset.

The remaining 7 groupings of comments all originate from the HITL Dataset. There were 420 ***Rejected AI Comments***. Human tutors edited 220[3] AI suggestions. We group the suggestions before editing as ***Pre-Edited AI Comments***. The corresponding 220 comment texts *after* tutor editing become their own feedback slice: ***Edited AI Comments***[4]. These latter two feedback slices can inform upon the value of human editing of AI suggestions. Tutors directly accepted 762 AI suggestions; we group these suggestions as ***Accepted AI Comments***.

The combination of rejected AI suggestions, pre-edited AI suggestions, and accepted AI suggestions create yet another feedback slice of 1,402 suggestions: ***All Pure AI Comments***. This feedback slice represents the set of comments that *would* have been surfaced to students without HITL intervention. Comparing this group against human only comments can illuminate differences in AI and human capability. However, to engage with AI responsibly, these AI-generated suggestions were never surfaced to students without human oversight. Nevertheless, we do report quantitative and qualitative comparisons against this group for the sake of completion.

We also include a smaller feedback slice of the 982 AI suggestions tutors chose to utilize: ***Final AI Interacted Comments (Accepted, Edited)***. This feedback slice only includes the comments students received as a result of a tutor accepting or editing an AI suggestion.

Lastly, we aggregate the ***Final HITL Comments (Human, Accepted, Edited)***. These 1,967 comments are the full set of comments surfaced to students after HITL intervention. Comparing this group to human-only comments (no AI access) will provide insights on the value of HITL tooling in terms of its impact on human ability; comparing against the set of all pure AI comments informs upon the utilization of AI automation.

---

[3] 11 suggestions originally stored as “edited” had to be removed since they also stored a previous “rejected” interaction. The true interaction could not be reconciled.

[4] 8 suggestions originally stored as “accepted” showed behaviors of being automatically “rephrased” by a downstream rephrase tool. We post-processed these suggestions as “pre-edited” and “edited” counterparts.

Automatic Comment Scoring

Paper has developed three generative, LLM-based[5], binary text classifiers to automatically score in-line written corrective feedback comments. In unison, these scorers cover three rubric dimensions: *encouraging*, *inquiry-based*, and *specific*. These dimensions are not a comprehensive rubric for written corrective feedback quality. Nonetheless, they provide a deeper understanding of the desirable, qualitative features a Review Center comment on the Paper platform should contain.

Given a single comment as input, each classifier generates a “yes” or “no” as output.[6] This “yes” or “no” response indicates the answer to a “key question” for each rubric dimension. These key questions as used by domain experts to label ground truth data are listed below. Domain experts also had access to examples and extended descriptions of each rubric dimension. We developed[7] each classifier using this labeled ground truth data and provide each classifier’s weighted and macro F1-score[8] performance on test data below in parentheses accordingly.

- *Encouraging* (.87; .67): Does the comment employ an encouraging and supportive tone?
- *Inquiry-Based* (.76; .75): Does the comment use inquiry-based questions to stimulate the student's thought on how to enhance or revise their work?
- *Specific* (.71; .71): Is the comment providing feedback which is specific to the student’s work and goes beyond offering generic advice?

We apply each scorer to all comments in both datasets. We aggregate the three “yes”/”no” responses for each comment. A response of “yes” indicates that a comment has met the qualifications for a certain rubric dimension; “no” indicates otherwise.

[5] The LLM used is GPT-4 (released on March 14, 2023) with a temperature of 1.5.
[6] A n value of 31 is used to call the model. A majority vote over the 31 “yes”/”no” responses and a self-consistency score between 0-1 is returned for classification.
[7] Detailed classifier development is out of scope for the present report.
[8] F1-score is reported on a 0-1 scale where a score of 1 indicates perfect accuracy. Weighted F1-score takes the size of the true number of “yes” and “no” labels into account; macro f1-score does not.

Analyses Conducted

We conduct a quantitative and qualitative analysis on the prevalence of *encouraging*, *inquiry-based*, and *specific* qualities in human-written, AI-generated, and notably, HITL written corrective feedback.

First, we present a quantitative analysis. We begin by examining the proportion of *encouraging*, *inquiry-based*, and *specific* comments between our 9 experimental groups of feedback slices. We report results in the section, *Across Comments*. To determine if proportions are significantly different between a pair of feedback slices, we conduct a single 2x2 Chi-Square Test of Independence for each pair of feedback slices for each rubric dimension (in total, 3 sets of 36 unique tests). Each Chi-Square Test compares the observed counts of “yes” and “no” scores for the pair feedback slices in question. On each set of 36 tests, we correct for False Discovery Rate (FDR) using the Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001). If an adjusted p-value is below the alpha level of 0.05, we conclude that the difference is statistically significant. We include tabulated raw counts and proportions of *encouraging*, *inquiry-based*, and *specific* comments in Appendix A. We place the tabulated differences in rubric dimension proportion and indicate significant results in Appendix B.

Next, we examine the average rate of *encouraging*, *inquiry-based*, and *specific* comments per submission. We compare the 200 submissions surfaced to students after HITL intervention and the 200 submissions from the Human Only Dataset. We also compare HITL submissions to the pure AI submissions (that were not surfaced to students). We report results in the section, *Across Submissions.* We use two-sample t-tests to determine if the average rate of each rubric dimension per submission is significantly different between feedback approaches. A significant difference indicates that the results shown are not likely to be due to chance. If the returned p-value is below the alpha level of 0.05, we conclude that the difference is statistically significant.

Our quantitative analysis wraps up with an examination of the number of rubric dimensions met concurrently. We compare the average number of co-occurring rubric dimensions per comment between human-only and HITL approaches, between pre-edited and edited comments, and between tutors with and without access to the AI tool. We report results in the section, *Rubric Dimension Co-Occurrence.* We also conduct three 2x4 Chi-Square Tests

of Independence to determine if the distribution of rubric dimensions met per comment are significantly different between feedback slices. Each Chi-Square test compares the observed counts of comments meeting 0, 1, 2, or 3 rubric dimensions for the two selected feedback slices. Again, if the returned p-value is below the alpha level of 0.05, we conclude that the difference is statistically significant.

Finally, we present a qualitative analysis of our data. We take a fine-grained look at individual human-written, AI-generated, and HITL edited comments. Such an examination is useful because binary text-classifiers do not directly output a "level" of encouragement, basis in inquiry, or specificity[9]. Only by manually reviewing comments or developing a more granular scoring system can we capture these insights. At this stage, we use the former. A Paper Teaching and Learning specialist brings to light areas of highest concern, errors in model scoring, strengths and weaknesses of AI-generated comments, and exemplifies the nature of human oversight. This qualitative analysis can thus inform areas for future exploration and comment scorer development. Future qualitative analyses should incorporate perspectives from multiple domain experts.

[9] As previously mentioned, we output a self-consistency score between 0-1 based on our classification models' 31 generations. Future exploration should confirm if this self-consistency score is a valid representation of "how" *encouraging*, *inquiry-based*, or *specific* a comment is against ground truth.

Results

In the following sections, we begin by presenting a quantitative analysis of the prevalence of *encouraging*, *inquiry-based,* and *specific* qualities in written corrective feedback according to automated scorers. We compare this prevalence across comments, across submissions, and as it relates to the co-occurrence of the aforementioned rubric dimensions. Unless mentioned otherwise, results reported are statistically significant. We finish with a granular qualitative analysis.

Quantitative Analysis

Across Comments

Encouraging

12% more HITL comments demonstrate an *encouraging* tone than comments solely composed by human tutors without AI tool access (42% and 30% respectively). This suggests a positive impact of AI involvement. However, comments from tutors with AI tool access are already 6% more often *encouraging* than comments from tutors without access (35.9% and 30% respectively). This complicates the attribution of gains in *encouraging* tone solely to the AI tool.

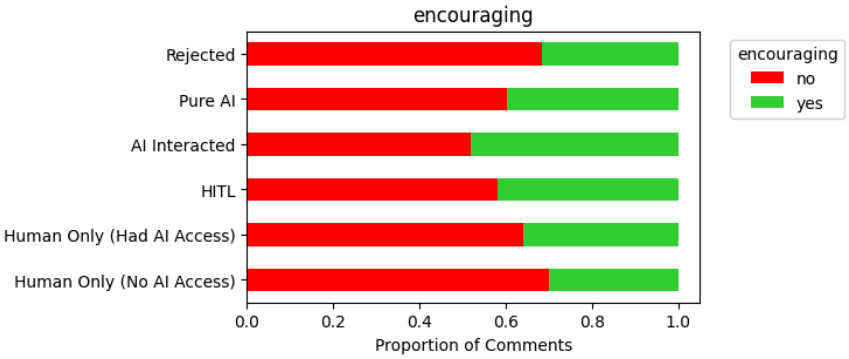

Figure 1. Proportion of encouraging written corrective feedback in human only comments, HITL comments, and AI suggestions. AI interacted (the combination of accepted and edited) suggestions demonstrate the largest proportion of encouraging comments. HITL feedback significantly surpasses human only feedback in encouraging comments. Pure AI feedback is significantly better than human only feedback without AI tool access but is not significantly better than HITL feedback.

Together, accepted and edited AI-generated suggestions are 12.1% and 18.1% more often *encouraging* than human-written comments by tutors with and without AI tool access. Interacting with AI suggestions promotes a higher rate of *encouraging* tone (48.1% of the time) than crafting a comment from scratch.

If a HITL approach is not taken, over half (60.3%) of the set of pure AI suggestions are not *encouraging*. However, this set of pure AI suggestions are 9.7% more often *encouraging* than comments solely composed by tutors without AI tool access. Even without human oversight, AI-generated comments provide a starting point to augment human capability.

Notably, rejected suggestions contain 10.3% less *encouraging* comments than the final set of HITL comments. Also, rejected suggestions are 16.4% less often *encouraging* than the group of AI suggestions that were edited and accepted by human tutors. They are 14.3% less often *encouraging* than just accepted suggestions, alone. Overall, 31.7% of rejected comments were *encouraging*, indicating that a suggestion with *encouraging* qualities does not preclude itself from rejection. For instance, suggestions that replicate previous feedback, include hallucinations, or have condescending tones can all be *encouraging*. Future explorations should examine the reason for rejection more closely.

Change After Editing	Proportion of Comments
Remains NOT Encouraging	.405
Remains Encouraging	.291
Becomes Encouraging	.264
Becomes NOT Encouraging	.041

Table 2. Proportion of changes in encouraging tone as a result of human comment editing.

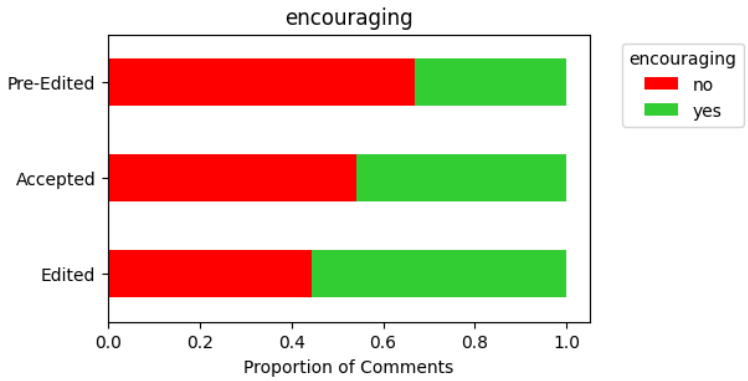

Figure 2. Proportion of encouraging tone in pre-edited, edited, and accepted AI suggestions. Editing an AI suggestion increases the chance for an encouraging tone.

Human oversight remains merited. Editing AI suggestions increases the rate of *encouraging* tone by 22.3% (from 33.2% to 55.5%). This surpasses the rate of encouragement in human-written comments with and without AI access by 19.5% and 25.5%, respectively. Edited AI suggestions are also 15.8% more often *encouraging* than pure AI suggestions lacking human oversight and 23.8% more often *encouraging* than rejected AI suggestions. While directly accepted AI suggestions are 12.7% more often *encouraging* than pre-edited ones, edited AI suggestions are 9.5% more often encouraging (albeit insignificantly) than the 45.9% of *encouraging* directly accepted ones. Looking even closer, 40.5% of pre-edited AI suggestions remain non-encouraging after editing, while 26.4% become *encouraging* after editing. Only 4.1% lose their *encouraging* tone. These results highlight the positive impact of human refinement in a HITL approach.

Inquiry-Based

The positive impact of HITL intervention is revealed when assessing comments on their rate of *inquiry-based* questioning. 9.7% more HITL comments are *inquiry-based* than those written by human tutors without AI tool access (40.8% and 31.1% respectively). The attribution of performance improvement solely to the AI tool is again complicated. Since, comments from tutors with AI tool access are already 6.8% more often *inquiry-based* than those from tutors without (37.9% and 31.1% respectively).

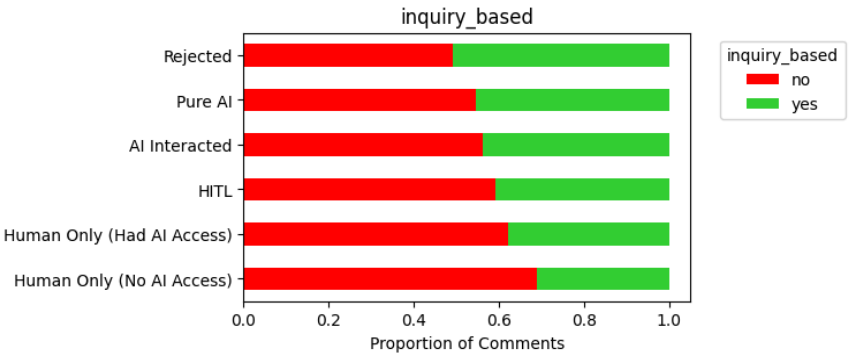

Figure 3. Proportion of inquiry-based written corrective feedback in human only comments, HITL comments, and AI suggestions. HITL feedback significantly surpasses human only (no AI access) feedback in inquiry-based comment amount. Pure AI feedback is significantly better than human only feedback but is not significantly better than HITL feedback. Rejected AI suggestions demonstrate the largest proportion of inquiry-based comments at a significantly greater rate than human only and HITL feedback.

Inquiry-based characteristics are demonstrated 12.6% more often when tutors directly accept or edit an AI suggestion as compared to when they write their own comments without AI tool access (43.7% and 31.1% respectively). To provide an *inquiry-based* experience more often, tutors should interact with AI suggestions.

As with *encouraging* tone, the full set of purely AI-generated suggestions give tutors a head start to success despite over half (54.5%) not displaying *inquiry-based* questioning. Comparing against comments constructed solely by humans with and without AI tool access, the set of pure AI suggestions are 7.6% and 14.4% more often *inquiry-based*, respectively. The 44.1% of *inquiry-based* accepted comments are 13% more often inquiry-based than comments authored by humans without AI access.

Rejected AI suggestions are significantly more often *inquiry-based* than human-only comments. The 50.7% of *inquiry-based* rejected suggestions are 12.8% and 19.6% more often *inquiry-based* than comments authored by humans with and without access to the AI tool, respectively. Rejected AI suggestions are also 9.9% more often *inquiry-based* than comments authored with a HITL approach. Thus, a comment with *inquiry-based* qualities does not preclude itself from rejection. Comments that replicate previous feedback, include hallucinations, have condescending tones, or even ask too many questions can all be *inquiry-based*. We again recommend that future explorations should examine the reason for rejection more closely.

Change After Editing	Proportion of Comments
Remains NOT Inquiry-Based	.450
Remains Inquiry-Based	.277
Becomes Inquiry-Based	.145
Becomes NOT Inquiry-Based	.127

Table 3. Proportion of changes in inquiry-based quality as a result of human comment editing.

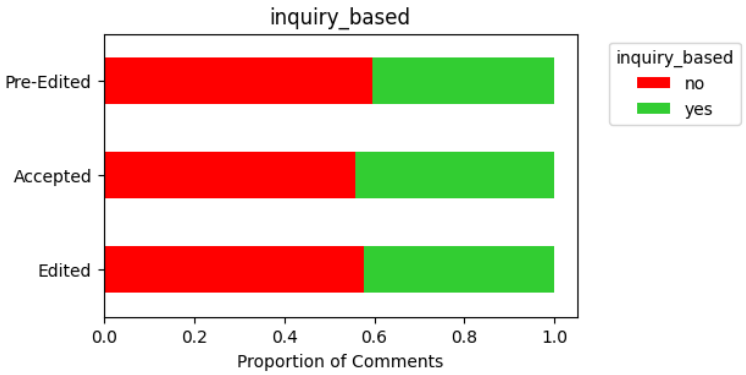

Figure 4. Proportion of inquiry-based feedback in pre-edited, edited, and accepted suggestions. Editing does not reveal a significant increase in inquiry-based suggestions.

Interestingly, human oversight through AI suggestion editing does not reveal any meaningful increase in *inquiry-based* comment amount (a statistically insignificant increase from 40.5% to 42.3%). Yet, while pre-edited suggestions are not significantly more often *inquiry-based* than comments written by humans without AI tool access, edited suggestions are, at a rate 11.2% higher. Digging deeper, we see fluctuations in the effects of human oversight. While 45% of pre-edited AI suggestions remain not *inquiry-based* after editing, 14.5% become *inquiry-based* after editing, and 12.7% lose their *inquiry-based* features. Crucially, not all comments need to be *inquiry-based*, such as comments intended for praise alone. Removing a suggestion's *inquiry-based* features can be desirable in such contexts. Future exploration could examine if such a context was taken into account for the 12.7% of suggestions which lost their *inquiry-based* quality. If so, human oversight adds this layer of needed vetting.

Specific

Unlike the previous two rubric dimensions, there is no significant difference between the ability of human tutors with and without AI tool access to comment with specificity (human-written comments are 45.4% and 45.3% *specific* respectively). Also, the 4.2% more HITL comments that are *specific* as compared to comments solely composed by tutors without AI tool access is not significantly larger. 49.5% of HITL comments are *specific*.

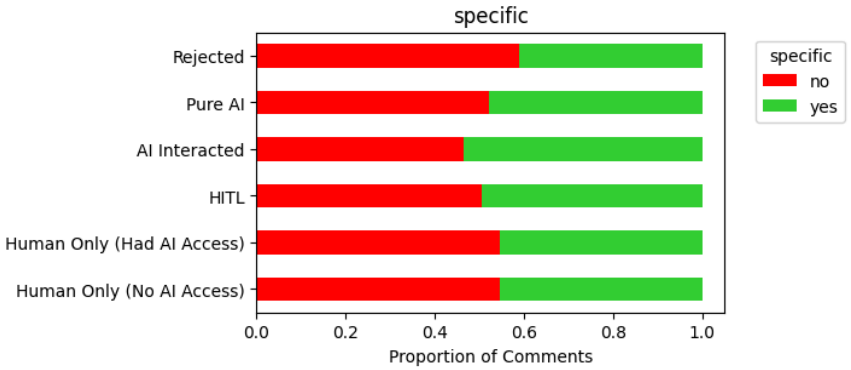

Figure 5. Proportion of specific written corrective feedback in human only comments, HITL comments, and AI suggestions. AI interacted (the combination of accepted and edited) suggestions demonstrate the largest proportion of specific comments. HITL feedback does not significantly surpass human only feedback in specific comments. Pure AI feedback is not significantly better than human only or HITL feedback.

However, interacting with AI suggestions via editing or directly accepting is 8.3% and 8.2% more likely to produce a *specific* comment than a tutor with and without AI tool access constructing their own comment, respectively. Results indicate that the option to interact with AI suggestions augment a human's ability to produce *specific* comments often (53.6% of the time).

Without HITL, just over half (52.1%) of the full set of AI-generated suggestions are not *specific*. The slight increase in *specific* comments as compared to human authored comments and slight decrease as compared to a HITL approach are both non-significant. AI generated suggestions alone may not give humans a head start in specificity.

Rejected AI suggestions do not contain significantly more *specific* comments than any feedback slice (41.2%). Notably, rejected suggestions are 8.3% less often *specific* than the final set of HITL comments and 12.4% less often *specific* than suggestions that were edited or

accepted by a human tutor. They are 9.7% less often *specific* than directly accepted suggestions, alone. Nonetheless, comments with *specific* qualities do not preclude themselves from rejection. We acknowledge once more that replicating previous feedback, hallucinatory content, condescending tones, or being so *specific* that a direct answer is provided can all produce *specific* comments. The value of insight on reason for rejection motivates a need for further exploration.

Change After Editing	Proportion of Comments
Remains Specific	.414
Remains NOT Specific	.282
Becomes Specific	.214
Becomes NOT Specific	.091

Table 4. Proportion of changes in specific quality as a result of human comment editing.

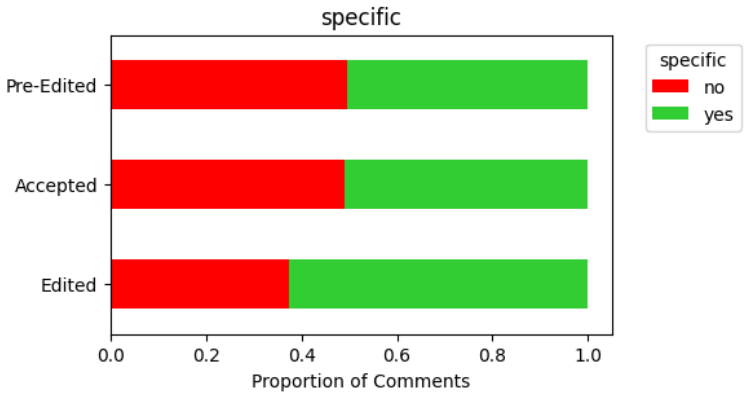

Figure 6. Proportion of specific feedback in pre-edited, edited, and accepted AI suggestions. Edited AI suggestions are non-significantly more often specific than accepted suggestions.

Editing insignificantly increases the rate of *specific* comments by 12.3% (from 50.5% to 62.7%). Yet, while pre-edited suggestions are not significantly more often *specific* than comments written by humans with and without AI tool access, edited suggestions are, at rates 17.4% and 17.3% higher, respectively. They are also 14.8% and 21.5% more often *specific* than the set of pure AI and rejected suggestions respectively. Lastly, edited AI suggestions are 11.8% more often *specific* than the 50.9% of *specific* directly accepted AI suggestions. A fine-grained look at these edits reveals that 28.2% of pre-edited AI suggestions remain non-specific after editing, while 21.4% become *specific* after editing, and 9.1% lose their specificity. These results again reinforce the positive impact of human refinement.

Across Submissions

Each student receives written corrective feedback in the format of a set of comments on a full submission (essay). In this section, we compare the average rate of rubric dimension alignment (*encouraging*, *inquiry-based*, and *specific*) between submissions reviewed with a human only approach and submissions reviewed with a HITL approach.

Within submissions reviewed by a human without access to the AI tool, an average of 31.7%, 31%, and 45.9% of comments per submission are *encouraging*, *inquiry-based*, and *specific* respectively. On the other hand, within submissions reviewed with a HITL approach, an average of 41.6%, 43.6%, and 51.3% of comments per submission are *encouraging*, *inquiry-based*, and *specific* respectively.

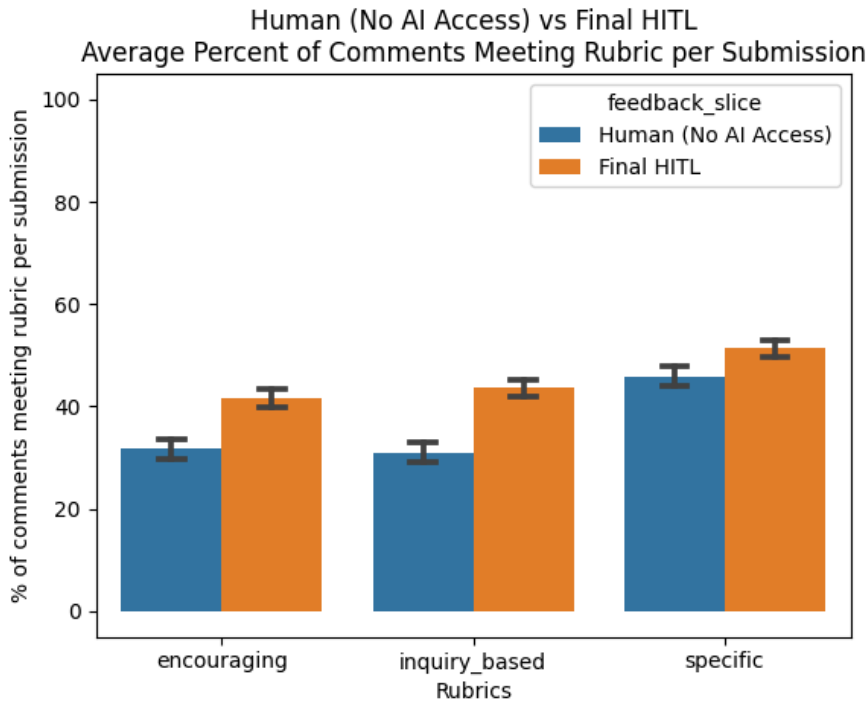

Figure 7. Grouped bar plot comparing the average percent of comments meeting a given rubric dimension per the two types of submissions surfaced to students. On average, submissions with HITL feedback contain significantly greater proportions of encouraging, inquiry-based, and specific comments than those with human only feedback. Error bars indicate a single standard error of the mean.

We find the percentage of comments meeting the *encouraging*, *inquiry-based*, and *specific* rubrics per submission are higher (+9.9%; +12.6%, +5.4%) when using a HITL approach as compared to tutors alone ($p<0.0002$; $p<0.0001$; $p<0.03$). When comparing submissions only reviewed by AI (that were not surfaced to students) against submissions reviewed with a HITL approach, we do not observe any significant differences for any of the three rubric dimensions ($p=0.155$; $p=0.931$; $p=0.185$). Despite this similarity, in the qualitative analysis below, we justify the benefits of human oversight.

		Encouraging	**Inquiry Based**	**Specific**
Human	**Mean**	31.7%	31%	45.9%
	Standard Dev.	27.8	26.3	26.4
HITL	**Mean**	41.6%	43.6%	51.3%
	Standard Dev	23.9	21.9	23.3
AI	**Mean**	38.4%	47.6%	48.2%
	Standard Dev	20.6	25.8	24

Table 5. Table expressing average value and variability in percentage of comments meeting each rubric dimension per submission between three types of submissions: human tutors without access to AI suggestions, the final set of HITL comments, and AI only reviewed submissions (that were not surfaced to students). Variability is considerable; HITL feedback is slightly less variable than human only feedback. AI only feedback is most variable.

Tutor only, HITL, and AI only reviewed submissions have considerable variability in the proportion of comments meeting a rubric. This indicates that the average proportion of comments meeting a rubric dimension are not always the case for individual submissions. If we compare the variability between groups by looking at standard deviation, we note that the amount of comments meeting a rubric dimension per submission in a HITL approach is slightly less variable than that of human only comments.

Rubric Dimension Co-Occurrence

Up until now, we have examined the occurrence of each rubric dimension independently of one another. In most cases, it is desirable for comments to meet all or a majority of rubric dimensions. Below, we compare the average number of co-occurring rubric dimensions between human-only and HITL approaches, between pre-edited and edited comments, and between tutors with and without access to the AI tool.

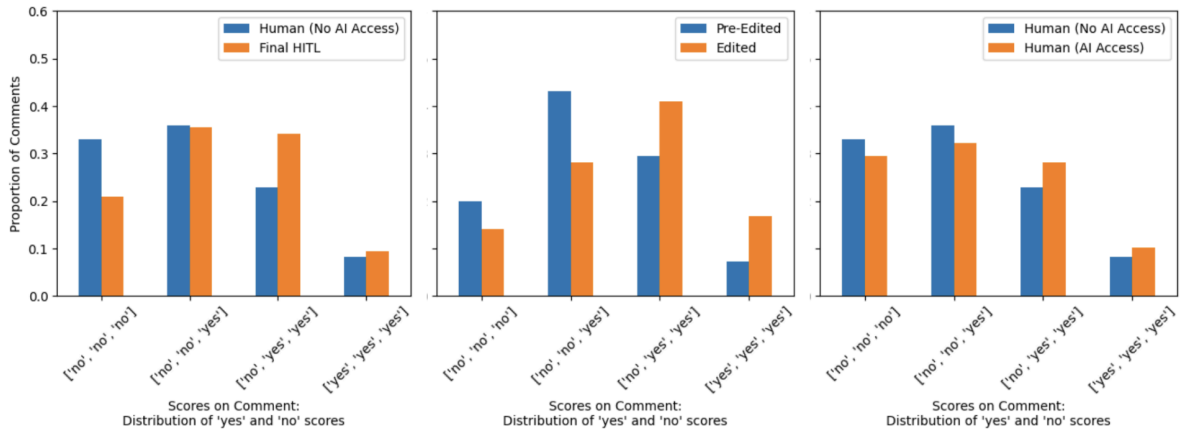

Figure 8. Grouped bar plots comparing the distributions of 'yes' and 'no' scores for three rubric dimensions on individual comments. HITL, editing, and human tutor differences positively influence the number of rubric dimensions met per comment.

Human (No AI Access) vs. HITL (Human, Accepted, Edited)

A HITL approach appears to reduce the proportion of comments not meeting any or only one of the three rubric dimensions while increasing the proportion of comments that meet at least two rubric dimensions. On average, a human without access to AI met 1.06 rubric dimensions per comment, while a HITL approach met 1.32 rubric dimensions per comment. Using a Chi-Square Test of Independence, we find that the distribution of rubrics met per comment is significantly different between a HITL and human only approach ($p<0.0001$). This indicates that a HITL approach is likely to increase the amount of rubric dimensions met by a comment. We note in a later section that human expertise was also significantly different between the two groups.

Pre-Edited vs. Edited

Tutors' edits of AI suggestions appear to increase the proportion of comments that meet two or three rubric dimensions. Editing also seems to reduce the proportion of comments not meeting any or only one of the three rubric dimensions. On average, a comment before editing met 1.24 rubric dimensions per comment, while comments after editing met 1.6 rubric dimensions per comment. Using a Chi-Square Test of Independence, we find that the distribution of rubrics met per comment is significantly different when choosing to edit a comment ($p<0.0001$). This indicates that tutor editing of AI suggested comments is likely to increase the amount of rubrics met by a comment.

Human (No AI Access) vs. Human (AI Access)

If a tutor had access to AI-generated suggestions, their human-only comments appear to have a higher likelihood to meet two or three rubric dimensions as compared to tutors without access. They also appear to demonstrate a lower likelihood to not meet any or only one of the three rubric dimensions as compared to tutors without access. On average, humans without access to the AI tool met 1.06 rubric dimensions per comment, while humans with access to the AI tool met 1.19 rubric dimensions per comment. Using a Chi-Square Test of Independence, we find that the distribution of rubrics met per comment is significantly different between both groups of human tutors ($p=0.0018$). This indicates that tutors selected to trial this HITL approach were more likely to already meet more rubrics on their own non-AI interacted comments. Hence, this introduces a potential confound for tutor ability.

Qualitative Analysis

Encouraging

AI-generated suggestions that score[10] as *encouraging* often provide exclusively positive feedback such as the following.

> *Your title is intriguing and makes me want to read more. Good job! A good title should be interesting and give the reader a hint about what the story will be about. Your title does both of these things very well.*

It is common for AI-generated suggestions that offer constructive feedback and score as *encouraging* to be implicitly *encouraging*. There is no overt validation for the student's efforts before highlighting an area for improvement.

> *What other details could you add to make the description of [Emily's]*[11] hometown even more vivid? Adding more details to your descriptions can help the reader visualize the scene more clearly and can make your story more engaging.*

Implicit encouragement is a divergence from Paper's tutor-facing pedagogical expectations, which emphasize that constructive feedback should be paired with compliments or validating statements to motivate students to improve. Consider the following constructive human-authored comments which score as *encouraging*:

> *Great work on your thesis so far! Now, let's make your thesis even stronger. You already have your main argument, which is great! Another step you can add is explaining why your main argument matters. This will help your readers care about and feel invested in this important topic! Here's what a full thesis statement might look like: -Students shouldn't have homework because it causes stress, results in less personal time, and doesn't contribute to their learning (what), with the consequence being that students cannot focus on their studies while in school, leading to bad grades (why). How can you add to your thesis so it explains why your ideas matter?*

> *Remember that last comment I left? Here, you did a great job of phrasing the sentence so that it is easy to read! "Elie and his father" is the subject, so "help each other persevere" makes sense. Does this help as*

[10] "Score" refers to the "yes"/"no" score provided by our binary text classifiers, not a ground truth score.
[11] Content surrounded by [ ]* indicates we minimally replaced content to maintain privacy of student work.

you edit the other sentence? (Remember a tutor can help in a Help session anytime if you need more support!)

The editing of AI suggestions demonstrates the ability of a human expert to align AI suggestions with pedagogical expectations for encouraging tone. Consider the following example of a pre-edited and edited suggestion within the HITL approach:

Pre-Edited AI-Generated Suggestion (Scored "No" for Encouraging)

Can you find a way to connect these two thoughts more smoothly? Connecting your thoughts more smoothly can help to improve the flow of your writing. In this case, breaking up the sentence into two separate thoughts and using a connecting word like 'made me realize' can help to make the sentence clearer and easier to read.

Edited Suggestion (Scored "Yes" for Encouraging)

You do a great job conveying the lesson your narrator learned! How might you connect these ideas more organically in your sentence? By doing so, you can create a more fluid narrative that will help simplify and clarify your point!

The pre-edited AI-generated suggestion is typical of the constructive feedback we see generated from our AI tool: it doesn't validate the student's efforts. This may come across to some students, particularly younger ones, as a bit dry or lacking in compassion. Human oversight ensures that AI-generated suggestions are simultaneously uplifting and constructive, without being overly prescriptive or harsh.

Inquiry-Based

In the Review Center, all comments (with the exception of exclusively positive feedback) must use an *inquiry-based* approach to help students think critically and independently. An *inquiry-based* approach asks meaningful questions that encourage exploration and independent problem-solving within an asynchronous context. Qualitatively, the *inquiry-based* approach is not applied consistently by our generative AI feedback tool, nor by human tutors.

For instance, the highest performing tutors use a mix of open- and close-ended questions to guide students to identify and revise an issue in their work.

> **Human-Written Comment - Open-ended Question**
> **(Scored "Yes" for Inquiry-Based)**
>
> *What was the narrator's first impression of the house or the lake? Did she enjoy being there? What was she feeling as soon as she arrived? Including these details here can give your readers a closer understanding of the character in your story :D*

Average tutors ask narrow, closed-ended questions that occasionally give the answer, which is not in line with an *inquiry-based* approach. The comment below is just barely *inquiry-based*:

> **Human-Written Comment - Close-ended question**
> **(Scored "Yes" for Inquiry-Based)**
>
> *This is a comma splice error because we're using a comma to join two complete sentences (without a conjunction). Is there another form of punctuation that would be more appropriate to use here?*

Close-ended questions become problematic when they get too specific or direct, as evidenced by the comment below. The *inquiry-based* classifier scored this comment incorrectly.

> **Human-Written Comment - Direct correction**
> **(Incorrectly Scored "Yes" for Inquiry-Based)**
>
> *Can you add a comma after the introductory phrase that I highlighted here? A comma should be placed after introductory phrases because they introduce the rest of the sentence. For example: ["As the cat licked his paw, he*

purred”] Do you see how [“he purred”]* could be a complete sentence on its own, but I chose to add [“as the cat licked his paw”]* as an introduction?*

AI-generated suggestions often follow similar patterns as average human comments, starting with the phrases "Can you..." or "How might you..." These questions sometimes imply a specific action for the student to take. The most problematic element of AI-generated suggestions like the one below are the direct suggestions that follow the question.

AI-Generated Suggestion - (Incorrectly Scored “Yes” for Inquiry-Based)

Can you think of a way to rephrase this sentence to make it more clear? In English, every sentence needs a subject and a verb to be complete. By starting the sentence with ['Getting on the bus,'] you make it clear that this action was the life-changing experience.*

The *inquiry-based* scorer is able to correctly flag some AI-generated suggestions that give direct answers, but the lack of precision emphasizes the value of using a HITL approach.

AI-Generated Suggestion - (Correctly Scored “No” for Inquiry-Based)

Can you think of a different transition word that might fit better in this context? 'Nevertheless' is usually used to introduce a contrasting idea or an unexpected outcome. In this context, a transition word like 'in the meantime' might be more appropriate because it shows that the character is doing something else while they wait to deal with the situation later.

Still, AI-generated suggestions can successfully implement the *inquiry-based* approach. The example below incorporates a healthy mixture of close and open-ended questions.

AI-Generated Suggestion - (Correctly Scored “Yes” for Inquiry-Based)

Could you delve a bit deeper into why a hammer and nails are considered male possessions? Why does this stereotype exist? Providing more context or explanation can help strengthen your point and make it more convincing!

Notably, the example above was among the set of rejected AI suggestions that scored as *inquiry-based*. This indicates that just because a suggestion is a stellar *inquiry-based*

suggestion is not reason enough to use it. Future explorations on rejection reasons should be conducted. Potential hypotheses include that rejected comments are condescending, replicate previous feedback, comment on low impact sentence level issues such as typos, or contain nonsensical advice.

For instance, the rejected comment below reads as positive feedback up until the last sentence, where a disjointed attempt at constructive questioning is made:

AI-Generated Suggestion

Great job starting your conclusion! This sentence effectively summarizes your argument and prepares the reader for your final thoughts. Starting your conclusion with a summary of your argument is a great way to remind the reader of your main points and prepare them for your final thoughts. Could you explain this even more professionally?

The findings above highlight the continued challenge of aligning AI-generated suggestions with requirements of the review workflow and tutor expectations.

Specific

Specificity personalizes a review and makes feedback actionable. Human tutors achieve this in two ways:

1. Pulling textual details from the student's work directly into comments to contextualize feedback.
2. Offering specific (not generic) advice or suggestions without direct corrections.

The following human-written comments that score as *specific* exemplify both methods, respectively.

> **Method 1.**
>
> *This is a great starting point to show how you fit the requirements of being a profiler but it could use some more detail. What are some real life examples you have of being a good communicator? Make sure to add some more detail to make this point stronger!*
>
> **Method 2.**
>
> *It looks like you have two complete thoughts separated by a comma. This is a type of run-on sentence called a comma splice. It means that you recognize that there are 2 complete thoughts (yay!), but you don't have them separated correctly.*
>
> *There are several ways you can correct this error.*
>
> *1) You can separate the sentences using a period, making two different sentences.*
> *2) Use a comma and coordinating conjunction (FANBOYS) to create a compound sentence.*
> *3) If the two sentences are closely related, you can use a semicolon to separate the complete thoughts.*
>
> *Which of these choices makes sense here?*

Specific textual details act as important context clues to help students understand the reasoning behind an issue and apply tutors' feedback to their work. *Specific* advice for rectifying comma splices is more actionable than broad, catchall tips about smooth sentence flow. AI-generated suggestions which score as *specific* commonly draw on

details from the student's work but issue generic writing advice about "flow" and "reader engagement":

> *How might you rephrase this sentence to avoid repeating the phrase 'has been through'? Repetition can sometimes make a sentence feel redundant and can disrupt the flow of your writing. By varying your language and sentence structure, you can make your writing more engaging and easier to read.*
>
> *Could you expand on why you've chosen these particular paths? Providing more detail about your choices helps your reader understand your motivations and makes your writing more engaging.*

Nonetheless, AI-generated suggestions can be successfully *specific* in detail and contextualized in advice. See this directly accepted AI-generated praise suggestion:

> *I love how you've given us a glimpse into the protagonist's life. It's a great way to build character and set the scene. Providing details about the character's living conditions and lifestyle helps to create a vivid image in the reader's mind and makes the character more relatable and interesting.*

Other AI-generated suggestions may be too *specific*—they engage with details from the student's work so directly that they give the student an explicit suggestion. This emphasizes the usefulness using both the dimensions of *specificity* and an *inquiry-based* approach to evaluate a suggestion. See an example below:

> *Can you think of a different transition word that might fit better in this context? 'Nevertheless' is usually used to introduce a contrasting idea or an unexpected outcome. In this context, a transition word like 'in the meantime' might be more appropriate because it shows that the character is doing something else while they wait to deal with the situation later.*

Many human-authored comments that do not score as *specific* are brief in length. Brief comments can still be *specific* in relation to the student's work. Future exploration could diagnose if our *specific* scorer often classifies such cases incorrectly.

> **Human Comments (Incorrectly Scored "No" for Specific)**

Check your spacing and capitalization here! How much space should be after a title like "Mr."? How do we usually capitalize people's names?

What are some kinds of shoes and clothes that you would buy?

Human Comment (Correctly Scored "No" for Specific)

Make sure to go back and proofread for missing punctuation!

Other short comments that do not score as *specific* rely on the context of student work.

Human Comment (Incorrectly Scored "No" for Specific)

This is very interesting, can you tell me more? This might be a great place for an example or how this impacted your life.

Sentence-level issues also seem to cause human-authored comments to not score *specific*. The comment below is detailed in its diagnosis of the issue and advice, even going as far as to include an example, but still incorrectly scores as not *specific*:

Just a small reminder that academic essays do not use contractions, as they sound too informal and conversational! Contractions are two words combined into one. For example, "don't" is the contraction of "do not." Are there any ways you can think to remember this writing tip? And how can you easily find contractions in this essay to spell them out?

As shown in the quantitative analysis, the merit of human oversight is demonstrated in AI suggestion editing. The same is true qualitatively.

Pre-Edited AI-Generated Suggestion

Could you consider rephrasing the phrase ['conveyed between the girl and the man] to make it clearer? Clear and concise phrasing helps the reader understand your point without any confusion.*

Edited Suggestion

Can you rephrase ["conveyed between the girl and the man"] to make it clearer? Clear and concise phrasing helps the reader understand your point without any confusion! Who is ["the girl"?]* Who is ["the man"?]*

While both the pre-edited and edited suggestions above score as *specific*, the addition of the final two questions contextualizes the advice. Editing helps indicate the characters in the sentence are unclear. The level of specificity is improved.

As a bonus, some edited suggestions which score as *specific* even demonstrate how tutors tweak details and context to align with the student's learner level:

> **Pre-edited AI-Generated Suggestion**
>
> *Could you clarify who the 'their' in 'their [achievements]* refers to? Is it [the Cherokee or the adults?]* It's important to make sure that pronouns like 'their' clearly refer to a specific noun. This helps your reader understand exactly what you mean.*
>
> **Edited Suggestion**
>
> *Hi, [John]*! What else could you tell us about the [Cherokee]*? Are they a group of people? If so, where are they from? When we add details, our ideas can really stand out :)*

Above, the tutor simplifies the content and language of the original suggestion to be more accessible for a grade 5 student. Cases like these suggest a need to consider how other factors play into and upon specificity in reviews. How *specific* details and advice are communicated matters as much as the presence of specificity itself.

Conclusion

Generative Large Language Models (LLMs) automate the production of text enriched with the frequent patterns exhibited in Natural Language. Leveraging this capability of LLMs while understanding their limitations is immediately relevant to numerous domains, particularly the educational technology sphere, as students encounter Artificial Intelligence (AI) more often in the classroom. In response to this call to action, at Paper we leverage generative AI to assist tutors in essay writing review.

Additionally, we utilize LLM-based scorers to automatically assess the efficacy of Human-in-the-Loop writing review feedback over select dimensions. Assessing writing review feedback manually is time-consuming. Automation allows large-scale, actionable feedback to be gathered and distributed downstream to stakeholders including tutors, AI writing tool developers, and academic communities at large.

In the present report, we demonstrated that LLMs can generate written corrective feedback with ease. We also quantified and showcased qualitatively that incorporating human oversight alongside AI-generated content extends the performance of human tutors. Written corrective feedback constructed via a HITL approach is more likely to be *encouraging, inquiry-based,* and *specific* than human only feedback. A HITL approach–particularly the editing of AI suggestions–combined with the differences in human tutors, aids in the co-occurrence of desirable feedback qualities.

Not only do many fine grained dimensions improve through the combined intelligence of AI and human experts, the pedagogical soundness of AI written corrective feedback is quality-checked. Human tutors have outstanding domain knowledge, unbounded context, and empathy; three qualities not accessible by ungrounded pattern generation alone. On the other hand, LLMs are fast and automate the production of text, giving tutors a jumpstart to their review process. To effectively integrate AI in student-facing products, a data-driven approach of “combined intelligence” respects the agency of educational professionals by equipping them with tools that enhance their performance, all without compromising on pedagogical quality.

References

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165-1188.

Chen, X., Wang, X., & Qu, Y. (2023). Constructing Ethical AI Based on the “Human-in-the-Loop” System. *Systems*, *11*(11), 548.

Klimova, B., Pikhart, M., & Kacetl, J. (2023). Ethical issues of the use of AI-driven mobile apps for education. Frontiers in Public Health, 10, 1118116.

Renz, A., & Vladova, G. (2021). Reinvigorating the discourse on human-centered artificial intelligence in educational technologies. *Technology Innovation Management Review*, *11*(5).

Sia, Dawn & Cheung (Nanyang Technological University), Yin Ling. (2017). Written corrective feedback in writing instruction: A qualitative synthesis of recent research. Issues in Language Studies. 6. 61-80. 10.33736/ils.478.2017.

Viola, M., de Queiroz, D., & Motz, R. (2023, June). Is the Human-in-the-Loop Concept Applied in Educational Recommender Systems?. In International Conference in Information Technology and Education (pp. 667-676). Singapore: Springer Nature Singapore.

Appendix A

Proportion of Encouraging Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
No	0.700	0.641	0.580	0.519	0.603	0.683	0.668	0.445	0.541
Yes	0.300	0.359	0.420	0.481	0.397	0.317	0.332	0.555	0.459

Table 6. Proportion of comments that score as “yes” and “no” for encouraging

Raw Counts of Encouraging Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
No	1254	631	1141	510	846	287	147	98	412
Yes	537	354	826	472	556	133	73	122	350

Table 7. Raw counts of comments that score as “yes” and “no” for encouraging

Proportion of Inquiry-Based Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
No	0.689	0.621	0.592	0.563	0.545	0.493	0.595	0.577	0.559
Yes	0.311	0.379	0.408	0.437	0.455	0.507	0.405	0.423	0.441

Table 8. Proportion of comments that score as “yes” and “no” for inquiry-based

Raw Counts of Inquiry-Based Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
No	1234	612	1165	553	764	207	131	127	426
Yes	557	373	802	429	638	213	89	93	336

Table 9. Raw counts of comments that score as “yes” and “no” for inquiry-based

Proportion of Specific Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
No	0.547	0.546	0.505	0.464	0.521	0.588	0.495	0.373	0.491
Yes	0.453	0.454	0.495	0.536	0.479	0.412	0.505	0.627	0.509

Table 10. Proportion of comments that score as “yes” and “no” for specific

Raw Counts of Specific Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
No	980	538	994	456	730	247	109	82	374
Yes	811	447	973	526	672	173	111	138	388

Table 11. Raw counts of comments that score as “yes” and “no” for specific

Appendix B

Difference in Proportion of Encouraging Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
Human (No AI Access)	0.00	0.06*	0.12***	0.181***	0.097***	0.017	0.032	0.255***	0.159***
Human (AI Access)	-0.06*	0.00	0.061*	0.121***	0.037	-0.043	-0.028	0.195***	0.1***
HITL	-0.12***	-0.061*	0.00	0.061*	-0.023	-0.103**	-0.088	0.135**	0.039
AI Interacted	-0.181***	-0.121***	-0.061*	0.00	-0.084***	-0.164***	-0.149***	0.074	-0.021
Pure AI	-0.097***	-0.037	0.023	0.084***	0.00	-0.08*	-0.065	0.158***	0.063*
Rejected	-0.017	0.043	0.103**	0.164***	0.08*	0.00	0.015	0.238***	0.143***
Pre-Edited	-0.032	0.028	0.088	0.149***	0.065	-0.015	0.00	0.223***	0.127**
Edited	-0.255***	-0.195***	-0.135**	-0.074	-0.158***	-0.238***	-0.223***	0.00	-0.095
Accepted	-0.159***	-0.1***	-0.039	0.021	-0.063*	-0.143***	-0.127**	0.095	0.00

Table 12. Difference in proportion of encouraging comments between feedback slices. Blue indicates the feedback slice on the X-axis is > the feedback slice on the Y-axis; red indicates otherwise.

*Statistical Significance: * indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$*

P-values are computed from 36 2x2 Chi Square Tests of Independence for the observed counts of "yes" and "no" scored encouraging comments for each unique feedback slice pair. P-values are corrected with the Benjamini-Yekutieli procedure for controlling the false discovery rate.

Difference in Proportion of Inquiry-Based Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
Human (No AI Access)	0.00	0.068*	0.097***	0.126***	0.144***	0.196***	0.094	0.112*	0.13***
Human (AI Access)	-0.068**	0.00	0.029	0.058	0.076**	0.128***	0.026	0.044	0.062
HITL	-0.097***	-0.029	0.00	0.029	0.047	0.099**	-0.003	0.015	0.033
AI Interacted	-0.126***	-0.058	-0.029	0.00	0.018	0.07	-0.032	-0.014	0.004
Pure AI	-0.144***	-0.076**	-0.047	-0.018	0.00	0.052	-0.051	-0.032	-0.014
Rejected	-0.196***	-0.128***	-0.099**	-0.07	-0.052	0.00	-0.103	-0.084	-0.066
Pre-Edited	-0.094	-0.026	0.003	0.032	0.051	0.103	0.00	0.018	0.036
Edited	-0.112*	-0.044	-0.015	0.014	0.032	0.084	-0.018	0.00	0.018
Accepted	-0.13***	-0.062	-0.033	-0.004	0.014	0.066	-0.036	-0.018	0.00

Table 13. Difference in proportion of inquiry-based comments between feedback slices. Blue indicates the feedback slice on the X-axis is > the feedback slice on the Y-axis; red indicates otherwise.

*Statistical Significance: * indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$*

P-values are computed from 36 2x2 Chi Square Tests of Independence for the observed counts of "yes" and "no" scored inquiry-based comments for each unique feedback slice pair. P-values are corrected with the Benjamini-Yekutieli procedure for controlling the false discovery rate.

Difference in Proportion of Specific Comments									
	Human (No AI Access)	**Human (AI Access)**	**HITL**	**AI Interacted**	**Pure AI**	**Rejected**	**Pre-Edited**	**Edited**	**Accepted**
Human (No AI Access)	0.00	0.001	0.042	0.083**	0.026	-0.041	0.052	0.174***	0.056
Human (AI Access)	-0.001	0.00	0.041	0.082**	0.026	-0.042	0.051	0.173***	0.055
HITL	-0.042	-0.041	0.00	0.041	-0.015	-0.083*	0.01	0.133**	0.015
AI Interacted	-0.083**	-0.082**	-0.041	0.00	-0.056	-0.124**	-0.031	0.092	-0.026
Pure AI	-0.026	-0.026	0.015	0.056	0.00	-0.067	0.025	0.148**	0.03
Rejected	0.041	0.042	0.083*	0.124**	0.067	0.00	0.093	0.215***	0.097*
Pre-Edited	-0.052	-0.051	-0.01	0.031	-0.025	-0.093	0.00	0.123	0.005
Edited	-0.174***	-0.173***	-0.133**	-0.092	-0.148**	-0.215***	-0.123	0.00	-0.118*
Accepted	-0.056	-0.055	-0.015	0.026	-0.03	-0.097*	-0.005	0.118*	0.00

Table 14. Difference in proportion of specific comments between feedback slices. Blue indicates the feedback slice on the X-axis is > the feedback slice on the Y-axis; red indicates otherwise.

*Statistical Significance: * indicates $p < 0.05$; ** indicates $p < 0.01$; *** indicates $p < 0.001$*

P-values are computed from 36 2x2 Chi Square Tests of Independence for the observed counts of "yes" and "no" scored specific comments for each unique feedback slice pair. P-values are corrected with the Benjamini-Yekutieli procedure for controlling the false discovery rate.