Using Large Language Models to augment (rather than replace) human feedback in higher education improves perceived feedback quality

Using Large Language Models to Augment (Rather Than Replace) Human Feedback in Higher Education Improves Perceived Feedback Quality

Thomas Schultze[1], Varun Suresh Kumar[2], Gary McKeown[1], Patrick A. O'Connor[1], Magdalena Rychlowska[1], and Kristina Šparemblek[1]

[1]School of Psychology, Queen's University Belfast, Belfast, United Kingdom

[2]School of Maths and Physics, Queen's University Belfast

Manuscript submitted for publication, draft version 2, 2024-04-12. This is a preprint, which has not yet undergone peer review. That means, the manuscript has not been finalized, might contain errors, and report information that has not yet been accepted or endorsed by the scientific community. If you cite this preprint or use the methods described here, please do so with caution.

Author Note

The study was preregistered. The preregistration, study materials, data, and analysis code can be obtained from https://osf.io/h8yxg/?view_only=2ebc16d82adf44c3aa06c55d6dcf1a1f.

Correspondence concerning this article should be addressed to Thomas Schultze, Queen's University Belfast, School of Psychology, 18-30 Malone Road, Belfast BT9 5BN, United Kingdom. E-mail: t.schultze@qub.ac.uk

Abstract

Formative feedback on assignments such as essays or theses is deemed necessary for students' academic development in higher education. However, providing high quality feedback can be time-intensive and challenging, and students frequently report dissatisfaction with feedback quality. Here we explore a possible solution, namely using large language models (LLMs) to augment feedback provided by instructors. One potential obstacle to using LLM-augmented feedback is algorithm aversion, which might lead students to deprecate LLM-augmented feedback. Therefore, we examined students' perceptions of human versus LLM-augmented feedback. In a pre-registered study, participants (N = 112) evaluated original human-generated versus LLM-augmented feedback on a previous assignment. Our results show evidence against algorithm aversion. Furthermore, participants rated the quality of LLM-augmented feedback substantially higher and strongly preferred it over the human-generated original. Our findings demonstrate the potential of LLMs to solve the persistent problem of low perceived feedback quality in higher education.

Keywords: AI; large language models; feedback; higher education; algorithm aversion

Using Large Language Models to Augment (Rather Than Replace) Human Feedback in Higher Education Improves Perceived Feedback Quality

1. Introduction

One of the pillars of higher education is formative assessment – informing students how to improve – with instructor feedback being a central component (Sadler, 1989). Formative feedback contributes substantially to students' progress (Hattie & Timperley, 2007; Kluger & DeNisi, 1996), particularly, when it includes explanations of the correct response and provides insights about what to improve and how (Heckler & Mikula, 2016; Lipnevich & Smith, 2009; Petrović et al., 2017). Formative feedback is particularly important for complex assignments such as essays or theses. However, providing quality formative feedback requires time. With growing student numbers and increasing academic workloads providing high-quality feedback is challenging, and a perceived lack of feedback quality among students is a recurring issue in higher education (Henderson et al., 2019; Mulliner & Tucker, 2017). For example, in the most recent UK's National Student Survey, students expressed dissatisfaction with the quality of the feedback they received (UK Office for Students, 2023), and this dissatisfaction is not a recent phenomenon (Price et al., 2011).

One way to address the problem of feedback quality in higher education is to use AI. AI and statistical systems have long been used in assessment, usually with a focus on automated grading. Page (1966) started the endeavour with Project Essay Grade, a regression-based grading system for high-school English essays. Later, systems such as Autotutor (Graesser et al., 2000) and the Intelligent Essay Assessor (Landauer et al., 2003) successfully used Latent Semantic Analysis to grade open responses based on their semantic similarity to ideal answers. Most work in automated essay scoring focused on style features like grammar, sentence structure, and punctuation. The development of techniques such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) allowed some improvement in content-based assessment. Automated essay scoring continued to improve with advancements in natural language processing. The first essay grading tools using large language models (LLMs) appeared in 2019, and these models can slightly outperform human raters (Rodriguez et al., 2019). In sum, AI-based systems are increasingly able to provide valid grades, especially when assignments focus on well-defined scientific concepts and sufficient training data exists (Ariely et al., 2023).

There are also attempts to use AI to provide formative feedback. Revision Assistant (Woods et al., 2017) uses predictive scoring to identify weak sentences in essays based on language, text organization, and support of arguments. Based on weaknesses identified, it then selects formative feedback from a set of statements pre-written by human experts. However, while good use of language is important in complex assignments such as essays, formative feedback on such assignments is ideally higher-level and content based. AI is unfortunately not (yet) able to provide valid evaluations or feedback on the classic essay format. For example, Dai et al. (2023) used an LLM to generate feedback on a student assignment and then assessed the validity of the LLM-generated feedback by comparing it to feedback from human experts. While the LLM-generated feedback outperformed human feedback in terms of readability, its validity was generally poor.

A possible solution to this problem is to use LLMs to augment rather than replace human feedback. We define LLM-augmented feedback as feedback generated by a human expert that is then processed by an LLM. Consider a human instructor who reads a student assignment and comments on its strengths and weaknesses. This expert then feeds those comments into an LLM, instructing it to make the text constructive and encouraging. Having an LLM augment the feedback might increase its perceived quality due to being more motivating. In addition, its perceived quality should be higher due to improved readability (Dai et al., 2023).

One potential obstacle to using LLM-augmented feedback is algorithm aversion, an irrational tendency to reject feedback generated using algorithms (Dietvorst et al., 2015). Algorithm aversion towards LLM-augmented feedback (compared to human feedback) might negatively impact students' learning trajectories. While there are no studies investigating students' acceptance of LLM-augmented – or LLM-generated – feedback, we can draw from research on advice taking. This research suggests that incorporating human judgements into algorithmic advice – similar to our idea of augmented feedback – moderates the degree of algorithm aversion, but there is still a controversy around the direction of this moderating effect (Jussupow et al., 2020). While some studies suggest that people derogate advisors who use algorithmic decision aids (Arkes et al., 2007; Shaffer et al., 2013; Wolf, 2014), others suggest that advice combining human and algorithmic judgment can reduce or even eliminate algorithm aversion (Himmelstein & Budescu, 2023; Palmeira & Spassova, 2015). In addition, some researchers argue that people prefer algorithmic to human advice, thus showing algorithm appreciation (Logg et al., 2019).

Given the mixed evidence concerning algorithm aversion and appreciation it is essential to assess how students perceive the quality of LLM-augmented feedback before we can consider employing it in higher education. Using LLM-augmented feedback might be counter-indicated if students show algorithm aversion toward it. If, however, there is evidence against algorithm aversion or for algorithm appreciation, then LLM-augmented feedback is viable tool in higher education. The present research aimed to test empirically whether students would show algorithm aversion or algorithm appreciation towards LLM-augmented feedback.

In a pre-registered study, we asked students to evaluate feedback on one of their previous assignments. We presented students human feedback they had received on that assignment (original feedback) along with an LLM-augmented alternative version. Since algorithm aversion or appreciation can only occur when students are aware that the alternative feedback was created using an LLM, we manipulated whether students were informed about or naïve to the use of an LLM when rating the feedback. Statistically, both algorithm aversion and algorithm appreciation should manifest as interactions between feedback type and awareness of the involvement of an LLM (awareness of the LLM's involvement should affect ratings of the augmented feedback while leaving ratings of the original feedback unchanged). This leads to two competing hypotheses.

H1a (algorithm aversion): When participants know how the augmented feedback was generated prior to rating it, their evaluation of the augmented feedback relative to the original feedback is more negative than that of participants who only learn about the generation process after the ratings.

H1b (algorithm appreciation): When participants know how the augmented feedback was generated prior to rating it, their evaluation of the augmented feedback relative to the original feedback is more positive than that of participants who only learn about the generation process after the ratings.

In addition to testing these interaction hypotheses, we explored how students perceived the overall quality of LLM-augmented feedback relative to its human-generated original (main effect of feedback type).

2. Method

2.1 Participants and Design

Participants were 112 second-year undergraduate psychology students at a UK university. Students were eligible to participate if they had been enrolled as first-year students in the same programme in the preceding academic year and had submitted a specific essay in that academic year (we used feedback students received on this essay as the stimulus materials). This limited the pool of potential participants to 193 people, which translates to a response rate of 58%. We did not collect demographic data, but as with most UK universities, the psychology student population mainly consists of UK nationals with notable proportions of foreign students from South Asia and Southeast Asia. The study used a 2 (*information about LLM*: after rating vs. before rating) × 2 (*feedback type*: original vs. augmented) mixed design with information about the LLM as a between-subjects variable and feedback type as a within-subjects variable. Participants were compensated with online vouchers worth £15.

2.2 Materials

Stimulus materials were developed using feedback participants had received from their instructors on an essay on individual differences submitted as part of the coursework of their first academic year. This feedback was provided by academic staff using three rubrics: strengths of the piece of work, what could be improved, and how to make improvements. Due to university regulations, feedback was standardized such that there was at least one point for each rubric, and the total number of points raised across rubrics was capped at nine. From this human-generated original feedback, we created an alternative version by feeding the comments into a localised instance of ChatGPT 3.5 (OpenAI, 2023). We instructed ChatGPT to make the original feedback constructive and encouraging.

2.3 Procedure

Participants were recruited for the online study via a mailing list, adverts in their lectures, and personalised email invitations containing a link to the study. Only eligible students could participate, and each person could take the survey only once. Participants were assigned to one of the two between-subjects conditions using an odd-even method. Every odd participant was assigned to the condition where information about the use of an LLM in the generation of the alternative feedback was provided before participants rated the feedback's quality. Even participants were assigned to the condition where we provided information about the LLM only after the quality ratings.

After reading the study information and providing informed consent, participants were directed to an instruction page. On this page, we informed participants that they would see the feedback they had received on their essay on individual differences in their first study year. We further told them that, in addition to the actual feedback they had received, we would show them an alternative version of that feedback. They would then be asked to read both versions of the feedback, rate their quality. In the condition providing the information about the LLM before ratings, the study instructions included the following paragraph:

*We would like to briefly explain how we created the alternative version of the feedback. To create it, we took the original feedback provided by your tutor and fed it into an AI, more specifically, a large language model (LLM). The LLM we used was ChatGPT, which you might be familiar with. We instructed the AI to take the original feedback and make it constructive and encouraging. The result is what we call **AI-augmented feedback**. AI-augmented feedback differs from AI-generated feedback in that it is based on human evaluation of your essay instead of an AI attempting to evaluate and provide feedback on its own.*

Next, participants were shown both the original and alternative feedback and were asked to rate the quality of each version. They then indicated which version of the feedback they preferred and explained their preference using an open-response format. Upon completing the ratings, participants were shown a debriefing page explaining the purpose of the study. For participants who received information about the LLM only after rating feedback quality, the debriefing included the paragraph shown above detailing how we generated the alternative feedback using an LLM. Following the debriefing, participants were thanked for taking the survey and received their online vouchers.

2.4 Measures

2.4.1 Feedback Quality

We measured feedback quality using the following four statements: "The feedback was clear and easy to understand", "The feedback provided specific suggestions for improvement", "Overall, I found the feedback helpful", and "I am satisfied with the quality of the feedback". Participants indicated their agreement with these statements using sliders ranging from 0 (complete disagreement) to 100 (complete agreement). We preregistered that we would aggregate the responses to the four items by averaging them if their internal consistency measured as Cronbach's α exceeded .70. This was the case for both the ratings of the original feedback (α = .90) and the alternative feedback (α = .80).

2.4.2 Feedback Preference

Our measure of feedback preference was a simple 2-alternative forced choice response. Participants could indicate either that they preferred the original or the alternative feedback. They were also asked to explain their preference using an open-response format.

3. Research Transparency Statement

Hypotheses, method, and analysis plan were preregistered on 12th December 2024 (https://osf.io/h8yxg/?view_only=092c4819de7f473c970292de30c12702). We deviated from the preregistration in one way: originally, we planned to stop collecting data at the end of November 2023. However, this would have left us with only 30 responses. Therefore, we decided to re-advertise our study using personalised emails and to continue collecting data until the end of December 2023. Study materials, data, and analysis scripts are publicly available (https://osf.io/h8yxg/?view_only=2ebc16d82adf44c3aa06c55d6dcf1a1f). All authors declare that they have no conflicts of interest. The study was supported by the Innovation in Teaching Fund of the Faculty of XXXX at XXXX. Our study was approved by the Research Ethics Committee of the Faculty of XXXX at XXXX on November 9th 2023 (reference number: EPS 23_394).

4. Results

We analysed the data using R version 4.3.1 (R Core Team, 2023) with the following packages: *BayesFactor* 0.9.12-4.6 (Morey & Rouder, 2023), *dplyr* 1.1.3 (Wickham, François, et al., 2023), *ggplot2* 3.4.4 (Wickham, 2016), *gridExtra 2*.3 (Auguie, 2017), *psych* 2.3.12 (Revelle, 2023), and *tidyr* 1.3.0 (Wickham, Vaughan, et al., 2023). We used Bayesian analyses with the Bayes factor (*BF*) as the inference criterion. As preregistered, we treated a $BF > 3$ as evidence for an effect, a $BF < 1/3$ as evidence against it, and *BF*s within this range as inconclusive.

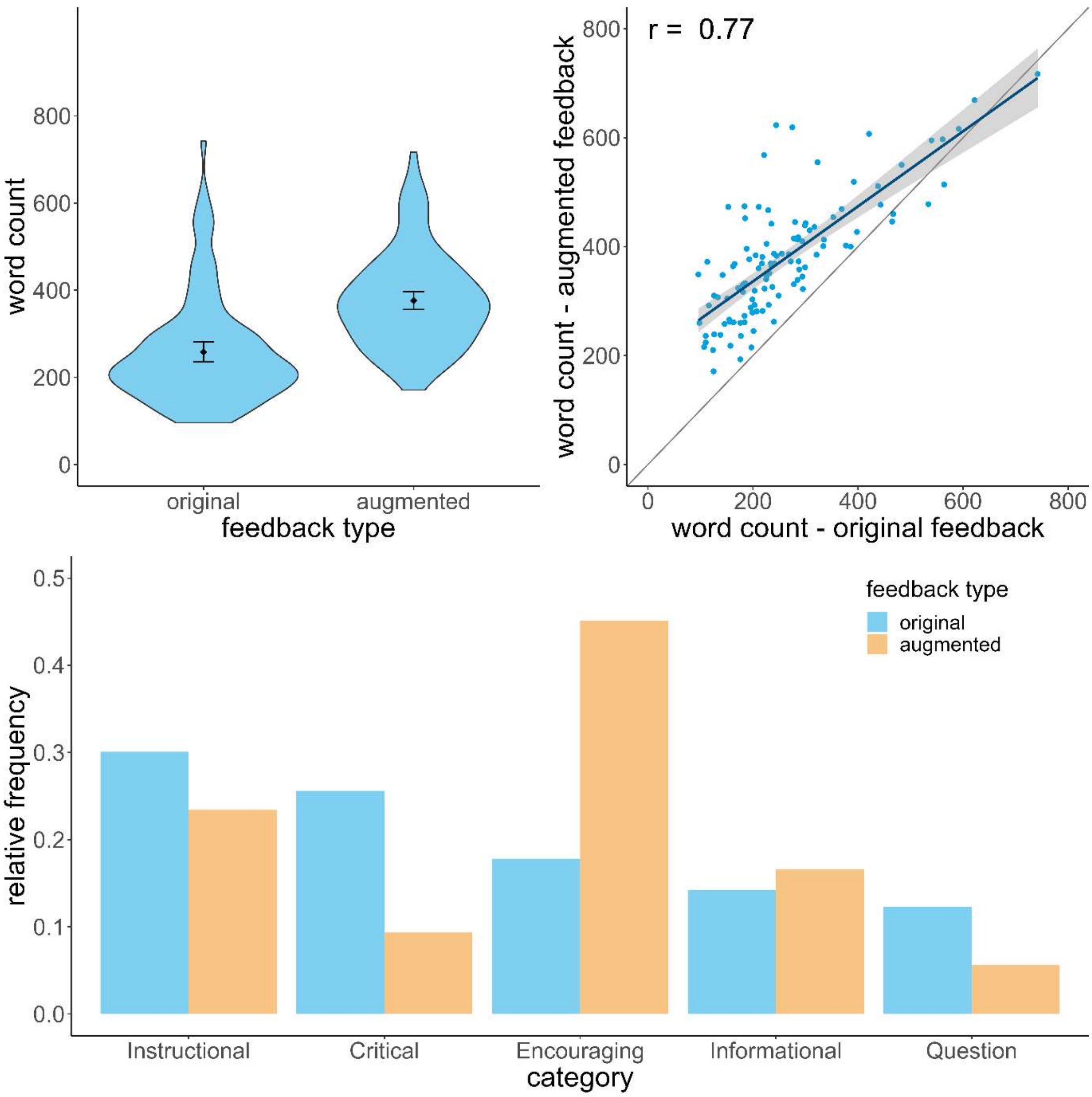
4.1 Preliminary Analyses

We compared the original and the LLM-augmented feedback using descriptive statistics. First, we looked at the length of the feedback, operationalized as its word count. As can be seen from Figure 1 (upper panels), the LLM-augmented feedback was, on average, about 100 words longer than the original feedback ($M = 376.39$, $SD = 110.19$ vs. $M = 258.35$, $SD = 122.87$), and this increase in feedback length was largely due to the LLM increasing the length of originally shorter feedback.

We then analysed the tonality of the feedback using another LLM, namely Facebook's bart-mnli model (Lewis et al., 2019). This model allows zero-shot classification of text, which means that it computes the probability of a piece of text to fall into certain user-defined categories. For the bart-mnli analysis, we split the feedback into individual sentences and then prompted the model to compute the probability of each sentence to fall into one of five categories. Categories were defined using the following labels: instructional, critical, encouraging, informational, and question. We then assigned to each sentence the label with the highest probability. The lower panel of Figure 1 shows the relative frequency of the five categories for each type of feedback. The most notable difference is that the tonality of the augmented feedback is less critical and more encouraging than the original feedback, whereas the relative frequencies of the remaining categories are relatively similar. This shows that our instruction to ChatGPT to make the original feedback constructive and encouraging worked as intended.

Figure 1

Comparison of the Original and the LLM-augmented Feedback



Note. The top left panel shows the word count by feedback type. Black diamonds and error bars represent means and 95% confidence intervals. The width of the violins represents the density of the data at the respective point of the y-axis. The top right panel shows how the length of the original feedback relates to the length of the augmented version. The line and bold line and grey band represent a linear regression line and its 95% confidence band. The bottom panel shows the results of a classification of each type of feedback using a bart-mnli model. Bars show the relative frequency of a sentence falling into a category.

4.2 Feedback Quality

The mean quality ratings by condition are shown in Figure 2. To test Hypotheses 1a and 1b, we ran a Bayesian mixed 2×2 ANOVA feedback quality. We used the default priors as implemented in the *BayesFactor* package for R (Morey & Rouder, 2023). In the Bayesian ANOVA, a null model, which only contains participants as random effects, is compared to four models: a model containing only the main effect of feedback type, a model containing only the main effect of information about the LLM, a model containing both main effects, and the full model with the main effects and their interaction. Each of these comparisons yields a *BF* (see Table 1).

Table 1

Results of the Bayesian ANOVA on feedback quality

Model	BF
ID + info about LLM	0.18
ID + feedback type	8.86×10^{7}
ID + info about LLM + feedback type	1.71×10^{7}
ID + info about LLM + feedback type + info about LLM × feedback type	4.33×10^{6}

Note. ID represents participants as random effects.

Since both hypotheses predict an interaction of feedback type and information about the LLM, we first tested for this interaction by comparing the full model (both main effects and the interaction effect) with the main-effects only model. The comparison yielded evidence against an interaction, $BF = 0.25$, that is, students' ratings were unaffected by the time at which they learned about how we generated the alternative feedback. In other words, there was neither algorithm aversion nor algorithm appreciation, which leads us to reject both H1a and H1b. We preregistered running simple effects analysis using Bayesian *t*-tests (scaling factor of $r = 0.50$ for the Cauchy priors on the effect size) only in case there was evidence for an interaction. However, for the sake of completeness, the results of these simple effects tests are included in Figure 2.

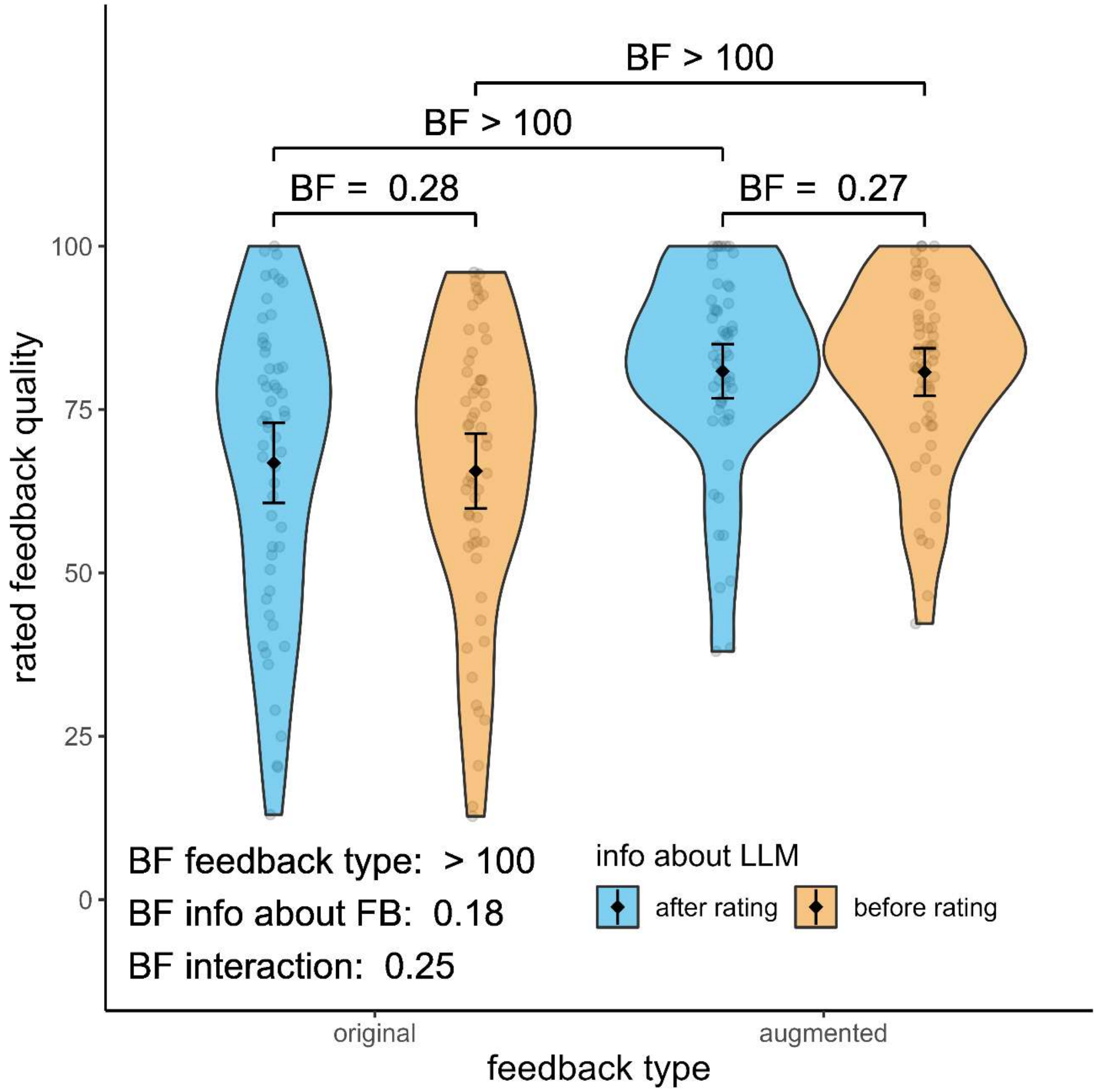
Having found evidence against an interaction effect, we next explored the main effects. Absent an interaction, the main effects can be tested by comparing the model containing only the main effect of interest with the null model. Feedback type had a strong impact on the quality ratings with the alternative feedback receiving much more favourable ratings than the human-generated original ($M = 80.80$, $SD = 14.50$ vs. $M = 66.20$, $SD =$

22.00). As can be seen in Table 1, the feedback type model not only fit the data better than the null model, but it also outperformed all other models (all *BF*s > 5.19). Information about the LLM had no effect on the quality ratings, $BF = 0.18$.

One simple explanation for LLM-augmented feedback receiving higher quality rating is that it is longer than the original feedback. To test this simple explanation, we ran a set of exploratory analyses. First, we tested within each feedback type whether word count was correlated with the quality ratings using the default Bayesian test of correlation against. We found evidence against a correlation of feedback length and quality for both the original feedback ($r = .01$, $BF = 0.22$) and the augmented feedback ($r = .07$, $BF = 0.29$). Furthermore, we tested whether word count predicted quality ratings over and above feedback type, information about the LLM, and their interaction. Since the Bayesian ANOVA does not allow including covariates, we ran two Bayesian regression models using default priors. The first model is the equivalent to the full model Bayesian ANOVA. It includes the two main effects and the interaction as fixed effects, and participants as random effect. The second model also includes word count as an additional predictor. Consistent with the absence of correlations within each feedback type, the model comparison yielded evidence against an effect of feedback length on its perceived quality, $BF = 0.29$.

Figure 2

Ratings of Feedback Quality by Feedback Type and Information About the LLM



Note. Black diamonds and error bars represent means and 95% confidence intervals. The width of the violins represents the density of the data at the respective point of the y-axis. The semi-transparent points represent individual quality ratings. Results of the Bayesian ANOVA are shown in the bottom left. Bayes factors for the simple effects analyses of feedback type and information about the LLM are displayed above the respective vertical brackets.

4.3 Feedback Preference

We ran an exploratory analysis on participants' feedback preferences to test whether they covaried systematically with information about the use of an LLM in the generation of the alternative feedback. To this end, we ran the default Bayesian test of non-independence implemented in the BayesFactor package on a 2 (*info about LLM*: after rating vs. before rating) × 2 (*preferred feedback*: original vs. augmented) contingency table. This test yielded evidence against non-independence, $BF = 0.20$. Among participants who received information about the LLM before rating it, 45 (79%) preferred the alternative feedback while only 12 (21%) indicated a preference for the original version. Numbers were comparable for participants who only learned about the use of an LLM after the ratings with 45 (82%) preferring the augmented feedback and 10 (18%) preferring the original. Finally, we tested whether participants preferred the alternative feedback above chance level using a default binomial test as implemented in the BayesFactor package. Consistent with the quality ratings, the test showed that participants systematically preferred the LLM-augmented feedback over the human-generated original, $\text{BF} = 2.76 \times 10^{8}$.

4.4 Open Responses

To gain more insights into students' reasons for preferring either the original or LLM-augmented feedback, we inspected their responses to the question asking them to explain their stated preference. For students who preferred the augmented feedback, three dominant themes emerged (a student's explanation could include multiple themes). The most frequent theme was that the alternative feedback was more positively phrased (76% of cases). It is important to note, here, that mentions of feedback positivity went beyond students preferring praise over criticism. In about half of the cases where students mentioned feedback positivity as a reason for preferring the augmented feedback, they added that – due to is more positive tone – they felt more motivated to engage with the feedback or more confident in their ability to improve. As one student put it:

"*Personally, regardless of the grade I would feel more confident in my ability to improve my next essays with the AI feedback because it is more uplifting and encouraging.*"

The second most prevalent reason for preferring the alternative feedback was that it was more detailed or clear on ways to improve (34% of cases), for example:

"*I found that the alternative feedback was a lot more constructive. I felt like the improvements section was elaborated a lot more than the first one.*"

The third recurring motive was better readability (27% of cases) as, for example, apparent from this comment: "*I feel that the alternative feedback is easier to follow and has a better flow while providing encouragement on the work and clearly describing what could be done better.*"

Among students who preferred the original instructor feedback, the three dominant themes were that the original feedback seemed more realistic (36% of cases), that students did not need ego-brushing (36% of cases), and that the original feedback was more concise (32% of cases). The following examples illustrate these recurring motives:

"*It isn't trying to protect your ego and make you feel better about yourself telling you it is super when there is [sic] actual improvements to be made. You trust it more as it's honest.*"

"*I rather it be more concise, the second one seems as though its [sic] trying harder to not offend me.*"

"*The original feels more genuine and doesn't try to curry favour by exclaiming that the work was impressive.*"

5. Discussion

We investigated how students in higher education perceive the quality of LLM-augmented formative feedback compared to human-generated feedback. We were particularly interested in whether students would show algorithm aversion or algorithm appreciation towards LLM-augmented feedback. Answering this question is essential before we can consider using AI to augmented feedback. The results of our study provide evidence against both algorithm aversion and algorithm appreciation. Students did not like the augmented feedback less or more just because they knew an AI was involved in its generation. The absence of algorithm aversion is critical because it is a license to explore the possibilities of LLM-augmented feedback in higher education further.

In addition, students rated the quality of the LLM-augmented feedback much higher than that of the original human feedback, and most of them preferred the augmented feedback. At first glance, this might seem trivial because we instructed ChatGPT to make the original feedback constructive and encouraging, and students might simply prefer feedback that praises them. However, students' open responses paint a more complete picture. The augmented feedback's more positive tone was associated with a greater reported motivation to engage with it and a bolstering effect on students' academic self-efficacy. Participants also felt that the AI-augmented feedback was easier to parse and more detailed regarding ways to improve. This finding is highly relevant considering the perceived lack of feedback quality in higher education mentioned above (Henderson et al., 2019; Mulliner & Tucker, 2017). Feedback can only be effective if students engage with it, and the greater perceived quality of LLM-augmented feedback along with its more encouraging tone might contribute to greater student engagement and better learning outcomes. Using LLM-augmented feedback might prove an effective remedy to the persistent problem of feedback quality in higher education.

We need to emphasise that LLM-augmented feedback is not a "one-size-fits-all" solution. A notable minority of students preferred the original feedback because they felt it to be more authentic, more concise, or because they felt that they did not need their teachers to brush their egos. Those students might be more likely to disengage from LLM-augmented than from human feedback. We cannot say, based on our data, what differentiates students who preferred the original feedback from those who preferred the augmented version. Understanding the individual differences underlying these preferences might allow using a tailored approach where student receive the type of feedback that is most effective for their improvement.

One interesting question is why students in our study neither showed algorithm appreciation nor algorithm aversion. One possibility is that these phenomena are largely restricted to the realm of judgement and decision-making. Another explanation, which we consider more realistic, is the involvement of the human element in AI-augmented feedback. In this regard, our findings are consistent with those of previous studies who found no evidence of algorithm aversion – or algorithm appreciation – toward hybrid advice (Lipnevich & Smith, 2009; Palmeira & Spassova, 2015). A recent study by (Himmelstein & Budescu, 2023) found algorithm aversion for political issues and appreciation for economic questions, but only for the a priori expectations of advice quality. When presented with the actual advice, there were no differences in advice taking. Perhaps, as a recent study insinuates, algorithm aversion is a question of hypotheticals whereas it disappears or even turns into algorithm appreciation when the advice is received (Logg & Schlund, 2024). Our findings, too, point toward the possibility that the perceived quality of feedback matters more than labels, at least in contexts where the recipient of the feedback can reasonably assess its quality or usefulness.

5.1 Limitations and Directions for Future Research

We would like to address three limitations of our study. First, our study used selfreport data from a single cohort of psychology undergraduates at a single university in the UK. Arguably, the characteristics of the students, the subject, the institution, or even the country or region a learning institution is placed in could influence students' acceptance of LLM-augmented feedback. Whether our findings generalise to other higher education contexts is an open empirical question. However, our study can serve as a blueprint for others who wish study the effectiveness of LLM-augmented feedback.

Second, while we found that students overwhelmingly preferred the AI-augmented feedback due to its more motivating tone, better readability, and more detailed instructions, we do not yet know whether LLM-augmented feedback also leads to better learning outcomes. Accordingly, future research should test the impact of LLM-augmented versus human-generated feedback on student engagement with the feedback as well as student progression and academic performance.

Finally, we cannot say, based on our results, to what extent the use of AI to augment human feedback could also be leveraged for greater efficiency. In our study, the original feedback comprised elaborate written responses to three rubrics. One interesting question is whether LLMs can generate high-quality feedback from more sparse human input, for example, by reducing that input to bullet points or – for recurring mistakes – predefined comments. Again, this is an empirical question that could be tested by comparing elaborate human feedback with LLM-augmented feedback generated from sparse human input instead. If the quality of LLM-augmented feedback based on sparse human input matches or even exceeds that of elaborate human feedback, then it could contribute to more sustainable academic workloads or free up valuable time for additional teaching opportunities.

References

Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine learning and hebrew NLP for automated assessment of open-ended questions in biology. *International Journal of Artificial Intelligence in Education*, *33*(1), 1–34. https://doi.org/10.1007/s40593-021-00283-x

Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *27*(2), 189–202. https://doi.org/10.1177/0272989X06297391

Auguie, B. (2017). *gridExtra: Miscellaneous functions for 'grid' graphics*. https://CRAN.R-project.org/package=gridExtra

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323–325. https://doi.org/10.1109/ICALT58122.2023.00100

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*, 114–126. https://doi.org/10.1037/xge0000033

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group, T. R. G., & Person, N. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*(2), 129–147. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT129

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heckler, A. F., & Mikula, B. D. (2016). Factors affecting learning of vector math from computer-based practice: Feedback complexity and prior knowledge. *Physical Review Physics Education Research*, *12*(1), 010134. https://doi.org/10.1103/PhysRevPhysEducRes.12.010134

Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education*, *44*(8), 1237–1252. https://doi.org/10.1080/02602938.2019.1599815

Himmelstein, M., & Budescu, D. V. (2023). Preference for human or algorithmic forecasting advice does not predict if and how it is used. *Journal of Behavioral Decision Making*, *36*(1), e2285. https://doi.org/10.1002/bdm.2285

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *ECIS 2020 Research Papers*. https://aisel.aisnet.org/ecis2020_rp/168

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 295–308. https://doi.org/10.1080/0969594032000148154

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension* (arXiv:1910.13461). arXiv. https://doi.org/10.48550/arXiv.1910.13461

Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, *15*(4), 319–333. https://doi.org/10.1037/a0017841

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Logg, J., & Schlund, R. (2024). *A simple explanation reconciles "algorithm aversion" and "algorithm appreciation": Hypotheticals vs. real judgments* (SSRN Scholarly Paper 4687557). https://papers.ssrn.com/abstract=4687557

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Morey, R. D., & Rouder, J. N. (2023). *BayesFactor: Computation of Bayes factors for common designs*. https://CRAN.R-project.org/package=BayesFactor

Mulliner, E., & Tucker, M. (2017). Feedback on feedback practice: Perceptions of students and academics. *Assessment & Evaluation in Higher Education*, *42*(2), 266–288. https://doi.org/10.1080/02602938.2015.1103365

OpenAI. (2023). *OpenAI GPT-3.5*. https://www.openai.com

Page, E. B. (1966). The imminence of... Grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238–243.

Palmeira, M., & Spassova, G. (2015). Consumer reactions to professionals who use decision aids. *European Journal of Marketing*, *49*(3/4), 302–326. https://doi.org/10.1108/EJM-07-2013-0390

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Petrović, J., Pale, P., & Jeren, B. (2017). Online formative assessments in a digital signal processing course: Effects of feedback type and content difficulty on students learning achievements. *Education and Information Technologies*, *22*(6), 3047–3061. https://doi.org/10.1007/s10639-016-9571-0

Price, M., Handley, K., & Millar, J. (2011). Feedback: Focusing attention on engagement. *Studies in Higher Education*, *36*(8), 879–896. https://doi.org/10.1080/03075079.2010.483513

R Core Team. (2023). *R: A language and rnvironment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R., & Medow, M. A. (2013). Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *33*(1), 108–118. https://doi.org/10.1177/0272989X12453501

UK Office for Students. (2023). *National Student Survey 2023*. https://www.officeforstudents.org.uk/data-and-analysis/national-student-survey-data/student-characteristics-data/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wickham, H., Vaughan, D., & Girlich, M. (2023). *tidyr: Tidy messy data*. https://CRAN.R-project.org/package=tidyr

Wolf, J. R. (2014). Do IT students prefer doctors who use IT? *Computers in Human Behavior*, *35*, 287–294. https://doi.org/10.1016/j.chb.2014.03.020

Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). Formative essay feedback using predictive scoring models. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2071–2080. https://doi.org/10.1145/3097983.3098160