

Exploring Instructors' Views on Fine-Tuned Generative AI Feedback in Higher Education

Anastasia Olga (Olnancy) Tzirides, Gabriela C. Zapata, Patrick A. Bolger, Bill Cope, Mary Kalantzis, and Duane Searsmith

Forthcoming. *International Journal on E-learning (IJTL)*.

Abstract

This paper explores the integration of Generative Artificial Intelligence (GenAI) feedback into higher education. Specifically, it examines the views of 11 experienced instructors on fine-tuned GenAI formative feedback of student works in an online graduate program in the United States. The participants assessed sample GenAI reviews, and their perspectives were recorded through numerical, best-adjective, and open-ended surveys. The findings revealed pervasively positive views across the AI feedback. Numerical survey results showed that the feedback was generally deemed relevant, clear, actionable, useful, and comprehensive. The best-adjective survey further specified the nature of these views. Open-ended responses supported both findings, suggesting that GenAI feedback aligned well with course rubrics and provided actionable suggestions. Nevertheless, some limitations were identified, such as redundancy and how lengthy suggestions could overwhelm students. The study offers suggestions for the improvement of fine-tuned GenAI feedback to improve its effectiveness and enhance higher education students' learning experiences, especially in online settings.

Keywords: Fine-tuned Generative AI; Formative Feedback; Instructors' Views

Exploring Instructors' Views on Fine-Tuned Generative AI Feedback in Higher Education

The benefits of formative feedback on university students' written work have been widely reported (e.g., see reviews by Morris et al., 2021 and Pearson, 2022). Studies have shown that personalized instructor comments can help students understand their own strengths and weaknesses, improve their writing, and, in turn, foster independent learning. With the recent rise of Generative Artificial Intelligence (GenAI) models, new approaches to formative feedback have emerged with the potential to reshape writing instruction in higher education (Cope & Kalantzis, 2024; Tzirides et al., 2024; Zapata et al., 2024). However, before wholeheartedly adopting GenAI feedback tools, it is essential to thoroughly evaluate their pedagogical affordances and limitations to determine how they compare to traditional human feedback.

In this paper, we do so by investigating experienced university instructors' pedagogical assessment of fine-tuned, AI generated feedback. First, we discuss the key literature on human feedback. We then illustrate this with a sample of feedback on a university student's work carried out by a fine-tuned GenAI model. This is followed by our own study.

Characteristics of Human Feedback on Written Work

Analyses of university instructors' formative feedback on written work (e.g., Hyland & Hyland, 2001; Pearson, 2022) have revealed that it generally includes praise, criticism, and suggestions aimed at motivating students, thereby boosting their self-efficacy, and offering actionable guidance for improvement guided by specific pedagogical goals/outcomes (Connors & Lunsford, 1993; Straub, 1997). Comments are also usually mitigated through hedging, question forms, and paired-act patterns (where criticism is softened by praise generally within the same sentence) (Hyland & Hyland, 2001). These strategies have been

shown to help prevent negative emotional responses that might hinder their progress (Pearson, 2022).

Additionally, research has uncovered similar organization patterns in instructors' feedback. For example, Mirador's (2000) analysis of comments on postgraduate students' writing in the UK identified a common rhetorical structure. Feedback usually starts with general impressions and the recapitulation of ideas, followed by the articulation of strengths and weaknesses, then suggestions for improvement, and finally the overall judgment of the work. This "clinching pattern" ensures feedback is both comprehensive and goal-oriented, resulting in reflection and action (Holmeier et al., 2018).

Generative AI Feedback Compared with Human Feedback

Current research suggests that effective instructor feedback is personalized, actionable, and aligned with learning objectives, all of which are achieved through a specific pragmatic-content, an organizational pattern, and mitigating structures. In this section, we provide an example of a fine-tuned GenAI review of a student's work, generated using our research lab's tool, to illustrate how it plays out (see the *Methods* for tool details).


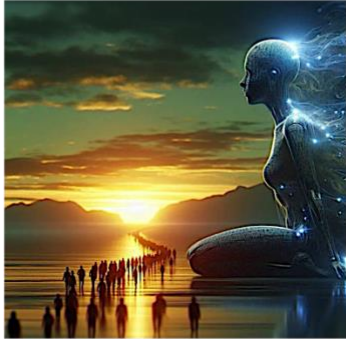
Figure 1 reveals that the content and organization are human-like. Specifically, the AI starts by offering an overview of the student's work and summarizes its main ideas, followed by the identification of strengths and weaknesses, then the provision of actionable suggestions, and finally an overall rating. The review also includes instances (highlighted in bold and italics) of hedging, paired-act patterns, and imperative forms. Such feedback features help establish a connection between the reviewer and the reviewee.

Figure 1

Presence of Characteristics of Human Feedback in GenAI Review of Student Work

Example of GenAI Review (Excerpt)

<p>(A) The essay provides a <i>comprehensive</i> examination of the implementation and implications of one-to-one technology in K-12 education. The writer <i>effectively communicates</i> the case by structuring the essay into <i>clear, logical</i> segments: Introduction, Theory, Practice, Analysis, and an intended Conclusion. The use of headings and subheadings, coupled with the integration of figures, <i>facilitates easy navigation and enhances the reader's understanding</i>. The essay <i>successfully integrates</i> figures to support its arguments, which <i>significantly enriches</i> the narrative. For instance, Figures 2 and 3 effectively illustrate the growth of individual learning devices and Chromebook shipments, respectively, providing empirical support for the claims made.</p> <p>(B) However, the essay could benefit from a more detailed discussion of each figure within the text to strengthen the connection between the visual data and the argumentation. The essay <i>maintains high textual coherence, with a smooth flow of ideas</i> from one section to another. The integration of media, such as figures and statistics, <i>effectively complements the text, although at times, the narrative could delve deeper into discussing the implications of these media on the overarching thesis</i>. For example, <i>while Figure 4 is mentioned, a more thorough analysis of its components and relevance to ubiquitous learning would fortify the argument</i>.</p> <p>(C) Constructive Suggestions for Revision:</p> <ol style="list-style-type: none"> 1. Elaboration on Media Items: <i>Ensure</i> that each figure and statistic is not only mentioned but also thoroughly analyzed in relation to the text. This will enhance the reader's understanding and the persuasive power of the essay. 2. Conclusion Development: The absence of a written conclusion leaves the essay feeling incomplete. A strong conclusion should synthesize the key findings, reiterate the significance of one-to-one technology in education, and perhaps suggest areas for future research or implementation strategies. 3. Addressing Counterarguments: While the essay mentions some drawbacks of one-to-one technology, it could benefit from a more detailed exploration of these challenges. Addressing and rebutting counterarguments would strengthen the essay's persuasiveness. 4. Expanding on Teacher Training: The section on professional development for educators using technology could be expanded. Given the importance of teacher readiness highlighted in the knowledgebase articles, further discussion on effective training models or examples of successful professional development programs would be valuable. <p>(D) Rating: The essay is rated a 4 for its very good presentation, structure, and use of media. It communicates the case effectively, with a clear and logical structure that is enhanced by the thoughtful integration of figures. However, to achieve a rating of 5, the essay would benefit from deeper analysis of embedded media, a written conclusion, and a more nuanced exploration of challenges and counterarguments related to one-to-one technology in education.</p>	<ul style="list-style-type: none"> • Presence of clinching pattern (Mirador, 2000): <ul style="list-style-type: none"> ◦ General impression, recapitulation, highlighting strengths (A) ◦ Calling attention to weaknesses (B) ◦ Offering suggestions for improvement (C) ◦ Overall judgement (D) • Mode (pragmatic intention of the feedback; Pearson, 2022): <ul style="list-style-type: none"> ◦ Advisory ◦ Praise ◦ Criticism ◦ Give information and need to (suggestions on resources and specific examples to undertake revisions) • Mitigating strategies (Pearson, 2022): <ul style="list-style-type: none"> ◦ Hedging ◦ Paired act patterns (positive softeners) • Direct dialogue with student (limited): <ul style="list-style-type: none"> ◦ Use of imperative forms
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Images created by students to represent their experiences with GenAI reviews (Tairides et al., 2024; Zapata et al., 2024)

In what follows, we investigate how experienced educators assessed GenAI reviews like the one presented in Figure 1. Our work focuses on the following research questions:

1. Do experienced university instructors feel that the GenAI feedback aligns with the pedagogical goals outlined in the assessment rubric?
2. Do experienced university instructors feel that the GenAI feedback effectively supports meaningful revisions?

Methods

This study examined experienced instructors' views on the quality of AI-generated feedback produced by a GenAI tool tailored for formative assessment of online graduate student work (MA, PhD, and Certificate) at a Midwestern US University.

Materials

The GenAI reviews were produced using OpenAI's GPT4 large language model via an application programming interface. To generate feedback on student writing, the GenAI tool was fed rubric criteria one-by-one, similar to how human reviewers operate. The rubric items and GenAI prompts were drawn from the *Learning by Design* framework (Cope & Kalantzis, 2010).¹ Additionally, the rubric-criteria prompts were supplemented by a separately specified knowledge repository in the form of a retrieval-augmented generation (RAG) database: The collective works of the program's faculty members and graduate students since 2019 (35 million words). The RAG database acts to push the AI towards generating feedback that is more domain relevant.

Participants

Eleven university instructors, with experience in higher education ranging from 3 to over 20 years, evaluated the GenAI feedback. The participants had a thorough understanding of the theory behind the rubric used by the AI, the GenAI's educational affordances and limitations, the online graduate-program curriculum, and the GenAI formative-feedback tool itself.

Data Collection

The participants assessed the GenAI feedback on three randomly selected graduate-student papers via two surveys: (1) A constrained-response survey with 7 numeric-rating and 6 best-adjective items, completed once per GenAI review for each student paper; and (2) an

¹ For a schematic view of the rubric, visit this link:
<https://drive.google.com/file/d/1DeciqZOZN8KR5EZNrxfk5meDTDeZXfoJ/view?usp=sharing>

open-ended survey with 19 questions, completed once per participant after reviewing the GenAI reviews.²

In the numeric-rating questions, participants ranked the GenAI review attributes of *relevance*, *clarity*, *actionability*, *usefulness*, *comprehensiveness*, and *objectivity*, along with their *overall satisfaction* in response to the following prompt: “On a scale of 1-5, where 1 is ‘Not at all’ and 5 is ‘Extremely’, how do you rate the feedback provided by the AI tool for each of these attributes?” In the best-adjective items, participants were asked to select from a list of 4-5 adjectives the one that best described the six attributes above (i.e., excluding *overall satisfaction*). These adjectives reflected those commonly associated with human feedback in previous literature and in our own work comparing GenAI and peer feedback from students’ perspective (e.g., Tzirides et al., 2023; Zapata et al., in press). The open-ended survey allowed instructors to elaborate on the same attributes. Due to potential ambiguity in interpretation, the item *objectivity* was removed from all our analyses. Since all instructors knew each other and some were more senior than others, all responses were completely anonymous at the outset.

Data Analysis

The numeric-rating and the best-adjective surveys were analyzed statistically with exact binomial and multinomial tests in the *R* programming language, version 4.4.2, “Pile of Leaves” (R Core Team, 2024).

The open-ended survey responses were analyzed thematically (Braun & Clarke, 2006). Specifically, we identified recurring themes that aligned with or expanded on trends in

² Due to space limitations, we cannot include the surveys in this paper. However, they can be accessed here:
Numerical questionnaire:

<https://drive.google.com/file/d/1rb1LXHvBdv7hL8WxoSovikY8q0jM2rfy/view?usp=sharing>

Qualitative survey:

<https://drive.google.com/file/d/1m4iLL942IRNW1vOaFdaibYXEXqPOyTKi/view?usp=sharing>

the numeric ratings. This mixed-methods analysis provides a deeper understanding of the participants' views.

Results

The results from all surveys were positive overall and provide complementary insights into the instructors' evaluations of GenAI feedback quality.

The numerical survey

Although the responses to the numerical survey might be conceived as interval data, there was little hope of assuming normal residuals. Analyses using ordinal, non-parametric approaches are problematic because of ties. Therefore, we transformed the ranks to two categories: *positive* and *non-positive* (negative and neutral). The positive scores were responses of 4 and 5 (where, recall, 5 corresponds to “extremely”). The non-positive scores were responses of 1-3 (where, recall, 1 corresponds to “not at all”). This allowed us to run an exact binomial test (2-sided) with respective expected probabilities under the null hypothesis of .6 and .4 respectively. That is, if participants responded randomly, then they would have chosen 1-3 60% of the time, and 4-5 40% of the time. Ultimately, this served as a test of whether participants were positive about the AI performance versus less than positive. To do this, we used the *binom.test* function from the *stats* package for R.

We also corrected for non-independent observations in the data. Recall that each instructor was allowed to respond to up to three of the AI reviews. But as noted above, we needed to keep the responses anonymous so that instructors felt free to respond honestly. For a binomial test, this is a violation of the assumption of independence (of observations). Thus, there was a non-trivial possibility of a Type I error if we maintained a rejection criterion of .05, all else being equal. Therefore, we reduced the alpha criterion to .01 to compensate. The results of these 2-tailed analyses are presented below in Table 1.

Table 1

Observed Counts and Associated p-values from Exact Binomial Tests Comparing Positive- vs. Non-positive Responses.

AI-review Attribute	Mean (SD)	Observed Counts (N=31)		p-values (all $ps < .01$)
		Non-positive (1-3) [†]	Positive (4-5) ^{††}	
<i>relevance</i>	4.48 (0.57)	1	30	.000000000022
<i>clarity</i>	4.32 (0.75)	5	26	.0000000081
<i>actionability</i>	4.32 (0.65)	3	28	.0000000075
<i>usefulness</i>	4.35 (0.75)	3	28	.0000000075
<i>comprehensiveness</i>	4.42 (0.67)	1	30	.000000000022
<i>overall</i>	4.39 (0.62)	2	29	.000000000050

[†] For all non-positive cases, the expected values under the null hypothesis are those closest to 60% of 31, or 18-19.

^{††} For all positive cases, the expected values under the null hypothesis are those closest to 40% of 31, or 12-13.

The null hypothesis of random responses across non-positive and positive values was rejected in all six cases (all $ps < .01$). Given the weight of responses being overwhelmingly positive, it is safe to assume that the instructors viewed the AI reviews to be positive across all attributes.

The best-adjective survey

The best-adjective survey provided further insights into the quality of AI feedback across the different attributes. To analyze whether and how many of these responses were unexpectedly high or low, we first ran exact multinomial tests as omnibus tests using the *EMT* package for *R* (Menzel, 2024). This is a generalization of the exact binomial test used above to more than two response categories. In this case, we had 4 or 5 response categories (adjectives), depending on the attribute. We followed up any significant effects with post-hoc, exact binomial tests (2-tailed) for each adjective (against the summed probabilities of the other adjectives in the group).

As above, we again reduced the alpha level to .01 from .05 to correct for hidden correlations among observations. Moreover, for any post-hoc tests that were justified, we

controlled for the familywise error rate associated with repeated testing by employing a Bonferroni adjustment of .01 divided by either 4 or 5 (depending on the number of adjective options for that AI-review attribute) to render new alpha criteria, respectively, of .0025 and .002. Table 2 below lists the attributes, their associated adjectives in the study, the observed counts for each adjective, and the results of the statistical analyses as *p*-values.

Table 2

Observed Counts and Associated p-values from Exact Multinomial (omnibus) and Exact Binomial (post-hoc) Tests Comparing Adjective Choice by Attribute. Statistical significance indicated with an asterisk and boldface.

Adjectives (under Attributes)	Observed Counts (N=31)	<i>p</i>-values
<i>relevance</i>[†]		.000279*
<i>unrelated</i>	0 (0%)	.000225*
<i>relevant</i>	11 (35%)	.211
<i>targeted</i>	14 (45%)	.0200
<i>precise</i>	6 (19%)	.541
<i>clarity</i>[†]		.0000413*
<i>confusing</i>	1 (3%)	.00282
<i>unclear</i>	2 (6%)	.0125
<i>understandable</i>	13 (42%)	.0373
<i>clear</i>	15 (48%)	.00563
<i>actionability</i>^{††}		.00127*
<i>vague</i>	0 (0%)	.00233
<i>general</i>	3 (10%)	.0182
<i>actionable</i>	12 (39%)	.0214
<i>specific</i>	9 (29%)	.258
<i>directive</i>	7 (23%)	.658
<i>usefulness</i>^{††}		.0000420*
<i>useless</i>	0 (0%)	.00233
<i>overly general</i>	3 (10%)	.182
<i>too specific</i>	0 (0%)	.00233
<i>helpful</i>	20 (65%)	.0000000875*
<i>valuable</i>	8 (26%)	.377
<i>comprehensiveness</i>^{††}		.0000420
<i>superficial</i>	0 (0%)	.00233
<i>basic</i>	1 (3%)	.0131
<i>adequate</i>	1 (3%)	.0131
<i>thorough</i>	17 (55%)	.0000189*
<i>exhaustive</i>	12 (39%)	.0214

[†] $\alpha=.0025$.

^{††} $\alpha=.002$.

The decision to apply such conservative null-hypothesis rejection criteria may have resulted in some Type II errors. This is apparent from the fact that although each attribute was statistically significant (all $ps < .01$), not all the associated post-hoc comparisons (within particular attributes) rendered significant results. For instance, the attributes of *clarity* and *actionability* showed no post-hoc differences when perhaps they should have.

But this conservative approach allows us to conclude that the instructors considered the AI reviews as not only not *unrelated* to the documents they reviewed, but also *helpful* and *thorough*. Again, given the conservative nature of our analysis, these are interpretations that we can consider minimally true, a very simple model, in other words: the AI reviews were not only (in a specific sense) helpful, thorough, and not unrelated to the student paper, but also (in a more general sense) clear and actionable.

The open-ended survey

The open-ended survey responses supported these findings. The instructors felt that the AI reviews were useful, as they aligned well with the course rubric by summarizing strengths and weaknesses across all criteria. For example, one instructor felt that “the AI reviews followed the provided rubrics closely and included both positive and negative aspects, highlighting areas for improvement and strengths clearly.” Another instructor echoed this sentiment, stating that the AI feedback was “highly aligned with [the rubrics]... every category [was] fully covered... [with] logical review responses that [were] potentially... actionable and focused.” Thus, they felt that the AI was effective in guiding students towards achieving their goals through contextualized suggestions across all criteria.

Instructors also highlighted the AI feedback’s ability to provide actionable suggestions tied directly to specific elements of students’ work. For instance, one respondent

believed “the AI review [had done] a good job compiling a list of improvement suggestions, preceded by explanations of what [was] working and what [was] not working [in the student’s essay], which [might have] helped [them] make sense of why those specific suggestions were made.” Additionally, instructors commended the AI feedback for being thorough and comprehensive. As one instructor observed, “[it] surpasses what a human or peer reviewer could typically offer... [with] insights and analysis that would take a considerable amount of time and effort for human or peer reviewers to generate.”

Despite these strengths, instructors also expressed limitations to the AI feedback. For example, they noted that the feedback occasionally lacked depth, failing to consider students experiences and motivations, which were sometimes critical for topic development. Furthermore, some instructors found the suggestions too general, lacking actionable detail. One stated that “the AI suggestions were sometimes too general, [lacking] the kind of specific detail that would help a student take the next step.”

Another limitation was the extensive, occasionally redundant nature of the feedback, which some instructors felt could overwhelm students. One described the feedback as “too wordy,” noting that “students [wouldn’t] spend much time reading almost 5,000 words in feedback because it’s time-consuming.” They felt that this verbosity, coupled with the occasional lack of concrete examples, could impede students from effectively identifying and prioritizing actionable steps.

Overall, while instructors praised the AI feedback for its alignment with rubric criteria and comprehensive analysis, they also highlighted the need to improve conciseness, personalized depth, and prioritization. They felt that addressing these areas could enhance the AI’s utility.

Discussion

The quantitative and qualitative analyses revealed that experienced instructors generally viewed the GenAI feedback positively. With respect to research question 1, for example (“Do experienced university instructors feel that the GenAI feedback aligns with the pedagogical goals outlined in the assessment rubric?”), the instructors seemed to believe that the AI’s feedback aligned well with the rubric’s intent, emphasizing the AI’s clear and structured guidance. Most instructors felt that the AI’s comments, suggestions for improvement, and ratings of students’ works incorporated the rubric criteria well and provided an accurate assessment of writing quality. This is consistent with the characteristics of effective human feedback highlighted by Hyland and Hyland (2001) and Pearson (2022), who argue that formative feedback should provide clear, structured comments, offering praise, constructive criticism, and actionable guidance reflective of instructional goals and expected outcomes to aid student improvement.

With respect to research question 2 however (“Do experienced university instructors feel that the GenAI feedback effectively supports meaningful revisions?”), the instructors’ views were more mixed. Although most felt that GenAI often provided helpful general advice, some felt it lacked specificity when translating suggestions into concrete actions, making the feedback less actionable. Additionally, instructors noted that GenAI’s recommendations were sometimes too exhaustive or redundant, potentially overwhelming students who might struggle to identify specific actions. This finding is important when considering the work of Connors and Lunsford (1993), Straub (1997), and Holmeier et al. (2018), who emphasize that effective feedback should include clear, actionable suggestions to foster meaningful revisions.

Although the example of the GenAI review in Figure 1 clearly exhibited characteristics and a structure like those found in human feedback, the results from this study

suggest that, in experienced instructors' views, AI feedback remains different from that offered by humans. These distinctions highlight certain opportunities for integrating GenAI feedback into higher education.

One example of an opportunity is that GenAI can quickly deliver structured, actionable feedback aligned with course rubrics and learning outcomes. This is particularly relevant to online and blended learning environments, where students might rely heavily on offline written feedback for guidance (Garrison & Vaughan, 2012; Means et al., 2013). In such settings, quick and comprehensive GenAI feedback can remove such delays, providing learners with an experience that emulates face-to-face feedback to a degree. However, to maximize effectiveness, the AI prompts (i.e., the rubric criteria) must be carefully crafted to generate actionable, detailed suggestions specific to each assignment.

Another opportunity is personalizing feedback. Instructors can develop prompts that guide the AI to scaffold student responses, encouraging reflective thinking about their work. This approach not only helps students to understand their strengths and weaknesses better, but also enhances the educational value of the feedback.

Additionally, it is essential to fine-tune the rubric prompts so that the AI output does not overwhelm students. By crafting prompts that ask for summarization and prioritization of key action items, instructors can ensure that feedback is concise. This is crucial in online settings, where students often review feedback independently and benefit from clear guidance on key priorities (Hrastinski, 2008).

In online and blended contexts moreover, where students often work independently, emotionally supportive feedback is critical to sustaining engagement and motivation (Kahu & Nelson, 2018). Incorporating mitigating strategies — such as hedging, paired-act patterns, and motivational phrases — can encourage students to persist in their efforts.

By refining prompts to include examples and explanatory comments that enhance learners' ability to interpret and apply feedback effectively, instructors can promote feedback literacy and independent learning. This approach ensures students can not only revise their work, but also critically assess feedback, fostering learner autonomy. In this way, GenAI feedback also becomes an integral part of the learning process, empowering students to take ownership of their development.

Along with pedagogical implications, this study also suggests areas for further research. Future research should investigate the adaptability of AI feedback tools not only across diverse educational and sociocultural contexts, but also within specific domains. Researchers could also explore how AI feedback fosters self-directed learning and long-term skill development. Additionally, the potential for GenAI to support metacognitive development through reflective prompts also warrants investigation.

At the policy level, although the integration of GenAI into higher education seems inevitable, there need to be guidelines that ensure ethical and equitable use. Institutions should develop frameworks that address issues such as data privacy, biases in AI-generated feedback, and accessibility for all students. Policies must emphasize transparency in how AI feedback is generated and used, ensuring that both students and instructors are well-informed about its capabilities and limitations. Additionally, training programs for faculty on using and interpreting AI feedback should be prioritized to bridge the gap between human expertise and technological affordances. As suggested by Cope and Kalantzis (2024), this could lead to reimagining instructional roles, with educators acting as facilitators and interpreters of AI-driven learning experiences.

Finally, the findings of this study underscore how important it is for practitioners, researchers, and policymakers to collaborate on exactly how to integrate GenAI into higher education. Insights from classroom applications should inform ongoing research and

technological development, while empirical findings should guide policy decisions that support innovation and equitable practices. Such an approach will ensure that AI tools are designed and implemented effectively and equitably, enriching both teaching and learning experiences.

As shown in this study, the integration of fine-tuned GenAI feedback into higher education offers significant potential to enhance the quality of formative assessment, especially in online settings. Alone, it can provide feedback both similar and at least complementary to that offered by human instructors. Nevertheless, there are still limitations. We propose that, by addressing them through targeted prompt refinement and fine-tuning, educators can ensure that AI feedback on student work is specific, actionable, and emotionally supportive. Combining AI-generated feedback with human insight can help provide personalized and comprehensive support to students, fostering their independent learning and encouraging meaningful revisions. Ultimately, the thoughtful use of GenAI, guided by well-crafted prompts and fine-tuning, can bridge the gap between automated and personalized feedback, resulting in an effective, richer learning experience for university students.

References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Connors, R. J., & Lunsford, A. A. (1993). Teachers' rhetorical comments on student papers. *College Composition and Communication*, 44(2), 200–223. <https://doi.org/10.2307/358839>
- Garrison, D. R., & Vaughan, N. D. (2012). *Blended learning in higher education: Framework, principles, and guidelines*. John Wiley & Sons.

Heiberger, R. M. (2024). *HH: Statistical analysis and data display: Heiberger and Holland*.

R package version 3.1-52.

Hyland, F., & Hyland, K. (2001). Sugaring the pill: Praise and criticism in written feedback.

Journal of Second Language Writing, 10(3), 185-212.

Holmeier, M., Grob, R., Nielsen, J. A., Rönnebeck, S., & Ropohl, M. (2018). Written teacher

feedback: Aspects of quality, benefits and challenges. In J. Dolin & R. Evans (Eds.),

Transforming assessment: Contributions from science education research (pp. 175-208). Springer.

Hrastinski, S. (2008). Asynchronous and synchronous e-learning. *EDUCAUSE Quarterly*,

31(4), 51-55. <https://er.educause.edu/-/media/files/article-downloads/eqm0848.pdf>

Hyland, F., & Hyland, K. (2001). Sugaring the pill: Praise and criticism in written feedback.

Journal of Second Language Writing, 10(3), 185-212. [https://doi.org/10.1016/S1060-3743\(01\)00038-8](https://doi.org/10.1016/S1060-3743(01)00038-8)

Kahu, E. R., & Nelson, K. (2018). Student engagement in the educational interface:

Understanding the mechanisms of student success. *Higher Education Research &*

Development, 37(1), 58-71. <https://doi.org/10.1080/07294360.2017.1344197>

Kalantzis, M., & Cope, B. (2024). Literacy in the time of artificial intelligence. *EdArXiv*.

<https://doi.org/10.35542/osf.io/es5kb>.

Kalantzis, M., & Cope, W. (2010). Learning by design. *E-Learning*, 7(3), 198-199.

Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The effectiveness of online and

blended learning: A meta-analysis of the empirical literature. *Teachers College*

Record, 115, 1-47. [https://www.sri.com/wp-](https://www.sri.com/wp-content/uploads/2021/12/effectiveness_of_online_and_blended_learning.pdf)

[content/uploads/2021/12/effectiveness_of_online_and_blended_learning.pdf](https://www.sri.com/wp-content/uploads/2021/12/effectiveness_of_online_and_blended_learning.pdf)

Mirador, J. F. (2000). A move analysis of written feedback in higher education. *RELC*

Journal, 31(1), 45-60. <https://doi.org/10.1177/003368820003100103>

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, 9(3), 1-26.

<https://doi.org/10.1002/rev3.3292>

Pearson, W. S. (2022). A typology of the characteristics of teachers' written feedback comments on second language writing. *Cogent Education*, 9, Article 2024937.

<https://doi.org/10.1080/2331186X.2021.2024937>

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Straub, R. (1997). Students' reactions to teacher comments: An exploratory study. *Research in the Teaching of English*, 31(1), 91–119. <https://doi.org/10.2307/40171265>

Tzirides, A. O., Zapata, G. C., Kastania, N. P., Saini, A. K., Castro, V., Abdul Rahman Ismael, S., You, Y., Afonso dos Santos, T., Sears Smith, D., O'Brien, C., Cope, B., & Kalantzis, M. (2024). Combining human and artificial intelligence for enhanced AI literacy in higher education. *Computers and Education Open*, 6, Article 100184.

<https://doi.org/10.1016/j.caeo.2024.100184>

Tzirides, A. O., Saini, A., Zapata, G., Sears Smith, D., Cope, B., Kalantzis, M., Castro, V., Kourkoulou, T., Jones, J., da Silva, R. A., Whiting, J., & Kastania, N. P. (2023). Generative AI: Implications and Applications for Education. *ArXiv*.

<https://arxiv.org/abs/2305.07605>

Zapata, G. C., Cope, B., Kalantzis, M., Tzirides, A. O., Saini, A., Sears Smith, D., Whiting, J., Kastania, N. P., Castro, V., Kourkoulou, T., Jones, J., & Abrantes da Silva, R. (In press). AI and peer reviews in higher education: Students' multimodal views on benefits, differences, and limitations. *Technology, Pedagogy and Education*.

Zapata, G. C., Saini, A. K., Tzirides, A., Cope, B., & Kalantzis, M. (2024). The role of AI feedback in university students' learning experiences: An exploration grounded in

Activity Theory. *Ubiquitous Learning: An International Journal*, 18(2).

<https://doi.org/10.18848/1835-9795/CGP/v18i02/1-30>