JUTLP
Journal of University Teaching & Learning Practice

Enhancing AI Feedback Systems in a Cyber-Social Learning Environment: Lessons Learned from a RAG Implementation Study

Christopher Hughes[a], Akash Saini[a], Mary Kalantzis[a], and Bill Cope[a]
[a] University of Illinois Urbana-Champaign, USA

Abstract

In this study, we compare graduate students' perceptions of AI-generated feedback before and after Retrieval-Augmented Generation (RAG) was added to a cyber-social learning environment. We used a non-equivalent cohort quasi-experimental design. We compared survey responses from Fall 2023 (n = 96) and Spring 2024 (n = 86). Survey items assessed perceptions of AI feedback in terms of *quality*, *usefulness*, and *actionability*. Results showed statistically significant between-cohort differences in perceived feedback quality (p = .003, d = 0.45) and usefulness (p < .001, d = 0.67), favouring the RAG cohort, while perceived actionability did not differ significantly between cohorts (p = .68, d = −0.06). Multiple regression analyses indicated that educational level was a positive predictor of perceived quality, usefulness, and actionability, whereas age and location were negative predictors of perceived actionability. These findings suggest that students in the RAG cohort perceived AI-generated feedback as more useful and better aligned with course expectations, particularly those at more advanced educational levels. However, the absence of improvement in perceived actionability suggests technical grounding alone may not suffice to help students translate feedback into revision decisions. This pattern underscores the importance of human support, feedback literacy development, and instructional scaffolding alongside AI-based feedback systems.

Editors

Section: Educational Technology
Editor in Chief: Joseph Crawford
Senior Editor: Rachel Fitzgerald

Publication

Submission: 10 Feb 2026
Revised: 15 Jul 2026
Accepted: 02 Oct 2026
Online First: 14 Oct 2026

Copyright © by the authors, in its year of first publication. This publication is an open access publication under the Creative Commons Attribution CC BY-ND 4.0 license.

Practitioner Notes

1. RAG implementation was associated with more positive student perceptions of the quality and usefulness of AI-generated feedback on student writing in graduate courses.
2. Educators implementing AI feedback tools should pair RAG-enhanced systems with human support, as RAG alone does not improve students' perceptions of feedback actionability.
3. Students pursuing advanced degrees tend to perceive AI-generated feedback more positively, suggesting that feedback literacy training should be targeted at less advanced learners.
4. Institutions adopting AI feedback should not assume that technological enhancements alone address the emotional and relational dimensions of feedback that drive student engagement.
5. Developers of AI feedback systems should consider adaptive mechanisms that adjust feedback complexity and specificity based on the learner's educational level and prior experience.

Keywords

Retrieval-Augmented Generation (RAG), Cyber-social learning, Generative AI, Feedback

Citation:

Hughes, C., Saini, A., Kalantzis, M., & Cope, B. (2026). Enhancing AI feedback systems: Lessons learned from a RAG implementation study. *Journal of University Teaching and Learning Practice, 23*(4). https://doi.org/10.53761/9vprqq24

Introduction

Generative AI (GenAI) has quickly become a promising tool for providing formative feedback on student writing in educational settings (Hughes et al., 2024). One reason for that promise is speed: these tools can return feedback quickly and in ways that appear aligned with task constraints (Tzirides et al., 2024). But the story is not uncomplicated. Persistent concerns remain around quality, usefulness, actionability, and the emotional flatness that can shape how students experience feedback in the first place (Saini et al., 2025). Large language models (LLMs), which power GenAI tools, are capable in interesting ways, but they also struggle in domain-specific and knowledge-intensive tasks (Gao et al., 2023). Among the best-known problems are hallucinations, particularly when responding to queries beyond their knowledge or requiring up-to-date information, reliance on non-authoritative or unreliable sources, and inaccuracies caused by confusion over domain-specific terminology (Zhang et al., 2023; Hadi et al., 2023).

One response to these limitations is Retrieval-Augmented Generation (RAG). When a user submits a query to a RAG-enabled system, the system first retrieves relevant material from external sources such as APIs, databases, or document repositories, often through embedding-based vector matching. That retrieved material is then combined with the user's prompt so the LLM can generate a response that is better grounded in relevant context (Gao et al., 2023). In practical terms, RAG matters because it can reduce inaccurate or weakly grounded output (Rangan & Yin, 2024). More broadly, it has become a key strategy for making generative AI more useful in real-world settings by improving currency, increasing developer control, and, in many cases, strengthening user trust (Balaguer et al., 2024; Fan et al., 2024).

This study examines whether the implementation of Retrieval-Augmented Generation (RAG) was associated with differences in graduate students' perceptions of AI-generated feedback in an online course environment. Importantly, the site of this study is a cyber-social learning environment (Tzirides et al., 2023), built intentionally on the principles of learning-by-design (Cope & Kalantzis, 2015). Learning by Design (LbyD) is a pedagogical framework that supports knowledge processes through experiencing, conceptualising, analysing, and applying, enabling learners to engage in active meaning-making across diverse and multimodal contexts. The learning platform studied is Common Ground Scholar (CGScholar). CGScholar is a cyber-social, "social knowledge" learning platform that supports collaborative engagement, peer-to-peer feedback, multimodal knowledge creation, mastery learning, publishing, and assess-as-you-go learning analytics. We share these details to help the reader understand that the site of this research project is a platform deliberately built on a pedagogical framework that encourages meaning making and recursive feedback that has gone through multiple iterations to foster dialogic learning.

The focus of the study is on perceived quality, usefulness, and actionability rather than on the objective quality of the feedback itself, students' revision uptake, or learning outcomes. This consideration is particularly important for AI-generated feedback. Regardless of the technical sophistication of an AI system, feedback is unlikely to support learning if students perceive it as low quality, unhelpful, or difficult to act upon. Consequently, learners' perceptions provide an important indicator of the educational value of AI-supported feedback. In this paper, we follow scholars such as Haughney et al. (2020) and Henderson et al. (2019) and use the construct *quality* to mean feedback that is clearly expressed, specific, and readily understandable to learners. We

follow Carless and Boud (2018) and Nicol and Macfarlane-Dick (2006) and use *useful* to mean feedback that supports learners' sense-making, self-monitoring, and self-regulation as they work to improve subsequent performance. We follow Wu and Schunn (2020) and Patchan et al. (2016) and use *actionable* to mean feedback that identifies what should be revised and provides guidance that can be translated into concrete changes in the learner's work.

We focus on learners' perceptions because feedback can improve learning only when students interpret and use it. Accordingly, understanding students' perceptions of feedback has become a central focus of feedback research (Van der Kleij & Lipnevich, 2021), while feedback literacy research further emphasises that learners must make sense of and act on feedback for it to improve learning (Carless & Boud, 2018).

There are, of course, many alternative ways to assess the value of feedback that warrant similar levels of investigation. Our goal is to formulate one piece of a larger puzzle. To examine differences in learners' perceptions between the non-RAG and RAG cohorts, we compared responses from two successive course cohorts. The Fall 2023 cohort (n = 96) completed the survey under the non-RAG condition, whereas the Spring 2024 cohort (n = 86) completed the same survey under the RAG condition. Statistical analyses revealed significant improvements in perceived quality and usefulness in the RAG cohort, with small to moderate effect sizes. However, no significant difference was observed in perceptions of actionability. Regression analysis, used to estimate the relationship between an outcome variable and one or more predictor variables, identified educational level as a significant predictor of positive perceptions across all dimensions. These findings suggest that the RAG implementation was associated with more positive student perceptions of feedback quality and usefulness, whereas perceptions of actionability did not improve and may depend on factors beyond technical grounding alone.

Research questions

RQ1. How do graduate students' perceptions of AI-generated feedback quality differ following the implementation of RAG?

RQ2. How do graduate students' perceptions of AI-generated feedback usefulness differ following the implementation of RAG?

RQ3. How do graduate students' perceptions of AI-generated feedback actionability differ following the implementation of RAG?

Theoretical Framework

This study is grounded in cyber-social learning theory, which reconceptualises the relationship between human and machine intelligence in educational settings (Cope & Kalantzis, 2024). In this framing, a cyber-social learning environment is one in which interaction, feedback, and knowledge production are distributed across people and digital systems, with AI operating as one part of a broader pedagogical ecology rather than as a replacement for human teaching. The theory matters here because it treats computational and human cognition as different but potentially complementary processes that, when brought together carefully, can produce new kinds of learning environments (Cope & Kalantzis, 2022a). It proposes five interconnected precepts: reflexive, collaborative, transpositional, multilinear, and non-scalar learning (Cope & Kalantzis, 2022b; Cope et al., 2021).

Three of those precepts are especially relevant to understanding how RAG may shape student perceptions of AI-generated feedback. The reflexive precept holds that learning is strengthened through frequent, low-stakes cycles of feedback, both human-to-human and human-to-machine, that allow learners to refine understanding and performance over time (Cope & Kalantzis, 2019). In a cyber-social learning environment, AI-generated feedback is one node in that broader reflexive cycle alongside peer and instructor review. The transpositional precept describes the movement between human intelligence and computer intelligence, where human meaning is rendered computationally actionable and computational output is made humanly meaningful (Cope & Kalantzis, 2024). RAG is a good example of that process because it brings human-authored knowledge artefacts, prior student work, instructor materials, and peer-reviewed literature into the generative process so that machine output is more clearly grounded in human meaning. The collaborative precept adds another layer by emphasising that knowledge construction is social and that learning environments work best when they draw on distributed perspectives across a community (Cope & Kalantzis, 2022b). That matters because actionability, the capacity to turn feedback into revision, may depend on this collaborative and dialogic dimension in ways that a technical intervention alone cannot fully address.

Taken together, these precepts lead to a clear theoretical expectation. If RAG strengthens the transpositional mechanism by grounding AI output in human-authored, context-specific materials, then students would be expected to perceive that feedback as higher in quality and greater in usefulness because the reflexive cycle becomes more accurate and more locally relevant. Actionability, however, may be more limited because acting on feedback depends not only on the quality of the feedback itself but also on learners' interpretation, agency, and participation within the broader collaborative learning environment (Tzirides et al., 2024). From the perspective of cyber-social learning theory, therefore, RAG is expected to strengthen some dimensions of learners' perceptions more readily than others. This theoretical perspective provides a useful lens for interpreting both the positive and the null findings in the present study.

Literature

The study of generative AI reviews sits within a much longer history of research on feedback in education. Hattie and Timperley (2007), for instance, distinguish between task-oriented feedback, which targets performance and learning outcomes, and process-oriented feedback, which helps develop autonomy, self-regulation, and metacognitive capacity. Both are concerned with closing the gap between current and desired performance, but they do so in somewhat different ways. Panadero and Lipnevich (2022), meanwhile, argue that higher education feedback models can be understood through four broad dimensions: content, function, presentation, and source, drawing on social-cognitive (Bandura, 1997) and constructivist (Vygotsky, 1978) traditions that emphasise reflection, guided interaction, and the development of capability through practice. More dialogic accounts of feedback move beyond the idea of feedback as a one-way transmission. Boud and Molloy (2013) and Carless and Boud (2018) frame feedback instead as an iterative process through which learners build evaluative judgment and participate in ongoing meaning-making around their work. In that sense, the rise of AI reviews can be understood as part of the broader shift described in Winstone and Carless's (2019) proactive feedback model, where students are positioned not as passive recipients of commentary but as active interpreters and users of evaluative information. Henderson et al. (2019) define feedback literacy as "the capacity to seek, understand, enact and produce evaluative information that leads to

improvement" (p. 1406), a formulation that is especially relevant here because it helps explain why the same feedback may be experienced differently by different learners. Fong and Schallert (2023) likewise emphasise the role of relational dynamics in how feedback is received and acted on, while Yeager et al. (2014) point to demographic differences as possible mediators of feedback receptivity.

GenAI Feedback

Empirical studies of AI-generated reviews show a familiar mix of promise and limitation. On the positive side, GenAI feedback is often viewed favourably and has been linked to improved academic performance (Pardo et al., 2019). AI systems are particularly attractive because they can provide both analytic and holistic evaluation in web-based environments (Shermis & Burstein, 2003), and students frequently describe such feedback as useful when revising drafts (Zhang & Hyland, 2018). There is also evidence that AI-supported learning can improve performance, as in work showing gains from ChatGPT-assisted learning (Pardos & Bhandari, 2024). In parallel, advances in natural language processing have expanded what AI tools can do in feedback-related contexts, including supporting peer review, identifying empathy, improving specificity, and incorporating background knowledge more effectively (Jia et al., 2022; Rietsche & Söllner, 2019). Researchers have also pointed to the consistency of AI-generated feedback across submissions (Mizumoto & Eguchi, 2023) and to its speed, which students often value highly (Seo et al., 2021). That timeliness matters because immediate feedback can be pedagogically powerful in both formative and summative contexts (Dikli, 2006; Hughes et al., 2024). Studies of ChatGPT-enabled tools extend this picture by suggesting that faster turnaround can be achieved while retaining relatively high feedback quality, although trained human evaluators remain stronger in accuracy and other important dimensions (Steiss et al., 2024).

At the same time, concerns about GenAI reviews remain substantial. Researchers have noted inaccuracies linked to biased training data (Harrer, 2023) and, in the case of ChatGPT, to insufficiently diverse training data (Dai et al., 2023). Building on Murphy's (2012) account of machine learning, Kirkpatrick et al. (2017) highlight a more general issue: AI models often struggle when tasks move beyond the domains or patterns represented in their training, which increases the likelihood of failure or distortion. Qualitative and mixed-methods work also suggests that even when students appreciate the utility of personalised AI feedback (Chan & Hu, 2023), they may experience it as less emotionally and socially engaging than human peer feedback (Al-Obaydi & Pikhart, 2025). In other words, AI feedback may be efficient and even helpful, while still feeling incomplete in ways that matter to learners.

Demographic variables further complicate the picture. Research on perceptions of AI and feedback shows patterns that are suggestive but inconsistent. Some studies report that male students express more positive attitudes toward educational AI (Stöhr et al., 2024), whereas others find no significant gender differences in overall GenAI perceptions or evaluations of AI-generated feedback (Nazaretsky et al., 2024). Age-related findings are similarly mixed: younger undergraduates have reported more positive attitudes toward GenAI (Klyshbekova & Forsberg, 2026), students aged 25–34 have viewed GenAI for academic purposes more positively than those aged 18–24 (Arowosegbe et al., 2024), and other work has found no significant age differences in perceived AI benefits or learning opportunities (Koohang et al., 2024). Findings concerning educational level are also mixed. Some studies associate higher educational

attainment or later years of study with more positive attitudes toward AI (Mukhamedkarimova & Umurkulova, 2025), whereas Henderson et al. (2019) suggest that perceptions of feedback may become less positive as students progress through higher education. Henderson et al. also report less positive feedback perceptions among domestic students than among international students. Taken together, these findings indicate that learner characteristics may shape how technical interventions are experienced, but they do not support a single consistent demographic pattern.

RAGs in Educational Contexts

Research on RAG in educational contexts has largely focused on whether it can address familiar GenAI problems such as verification, transparency, authenticity, and accountability. In broad terms, RAG is often presented as a way of making AI-generated responses more reliable by grounding them in traceable external materials (Elmessiry & Elmessiry, 2024). That promise is appealing in education, where trust, accuracy, and contextual relevance matter a great deal. At the same time, the literature does not present RAG as a magic fix. Researchers have pointed to ethical concerns such as privacy and bias, to technical and language limitations, and to the risk that users may become overly reliant on AI tools (Miladi et al., 2024; Rashid et al., 2024). Other practical concerns include the system's dependence on text-based content, difficulty handling images or equations, and the challenge of integrating diverse knowledge sources effectively (Dakshit, 2024). There is also evidence of low engagement and mixed perceptions of usefulness, with calls for more work on individual differences, functionality refinement, and long-term effects on feedback literacy (Thüs et al., 2024).

Even so, much of the literature is optimistic about what RAG can contribute. Broadly speaking, RAG is said to support more precise and contextually relevant responses that can be useful for personalised learning, automated content creation, and virtual tutoring (Rashid et al., 2024). Several studies report improvements in factual accuracy when answers are grounded in course materials, along with reductions in hallucination and stronger perceptions of trustworthiness (O'Dell & DeOrio, 2025). Other work describes RAG-based systems as virtual teaching assistants capable of answering student questions and generating teaching materials in more context-sensitive ways (Dakshit, 2024). Elsewhere, RAG has been shown to reduce hallucinations and produce more accurate, context-specific responses in higher education settings (Thüs et al., 2024). A RAG-augmented GPT-4 intervention, for example, was found to improve accuracy and reliability while reducing hallucinations, suggesting significant potential for more tailored and credible AI support in education (Miladi et al., 2024).

Research focused specifically on reviews of written assignments points in a similar direction. RAG has been shown to improve scoring accuracy by retrieving semantically similar examples from a vector database during inference and incorporating them into prompts (Wang & Ormerod, 2024). Other work reports that, across criteria such as context relevance, factual accuracy, completeness, logical consistency, and grammar, a RAG-enhanced ChatGPT tool produced detailed and consistent evaluations and was associated with improved final grade outcomes (Jauhiainen & Garagorry Guerra, 2024). At the same time, better scoring alignment does not automatically mean students will find feedback clear, motivating, or easy to act on. That distinction matters for this study. A system can be more technically grounded and still leave open the question of how learners actually experience the feedback it produces.

RAG may improve scoring accuracy by supplying the model with semantically similar, rubric-relevant examples or reference materials at inference time, thereby reducing reliance on generic response patterns alone. In practical terms, this can improve alignment between generated comments and assessment criteria. However, stronger scoring alignment does not necessarily guarantee that students will find the feedback clear, motivating, or easy to implement, which is why perceptual and behavioural dimensions must be distinguished. Comparatively little attention has been paid to whether these technical improvements (such as factual accuracy, hallucination reduction, retrieval performance, or scoring consistency) translate into educationally meaningful differences from the learner's perspective.

Hypotheses

H1: Students in the RAG cohort will report more positive perceptions of the quality of AI-generated feedback than students in the non-RAG cohort.
H2: Students in the RAG cohort will report more positive perceptions of the usefulness of AI-generated feedback than students in the non-RAG cohort.
H3: Students in the RAG cohort will report more positive perceptions of the actionability of AI-generated feedback than students in the non-RAG cohort.

Method

Study Design and Context

This study used a non-equivalent cohort comparison quasi-experimental design to examine whether the implementation of Retrieval-Augmented Generation (RAG) was associated with differences in students' perceptions of AI-generated feedback across two adjacent semesters. The Fall 2023 cohort served as the non-RAG comparison group (n = 96), and the Spring 2024 cohort served as the RAG group (n = 86). Both cohorts completed seven 8-week online graduate courses at a Midwestern public university in the United States using the same platform-based review workflow. In other words, the study was designed to capture what happened when a live instructional system changed under authentic teaching conditions rather than in a tightly controlled laboratory setting.

Introducing RAG into the AI feedback system affected the entire course environment, and withholding the upgraded system from some students during the same semester would have created unequal instructional conditions. Consequently, the intervention was evaluated using a quasi-experimental comparison of two successive course cohorts. Because the cohorts occurred naturally and participants were not randomly assigned, the design cannot rule out cohort effects, historical differences between semesters, or unmeasured differences in students' prior experience with AI and feedback. The findings should therefore be interpreted as evidence of cohort-level differences associated with the introduction of RAG rather than definitive evidence of causal effects.

To reduce those threats as much as possible, the instructional context was held as constant as it realistically could be across semesters. Both cohorts completed the same course, worked with identical learning objectives, followed the same assignment sequence and essay-writing workflow, used shared assessment rubrics, and participated in the same iterative feedback cycle involving AI-generated feedback, peer review, and instructor evaluation. The same course

platform and survey instrument were used in both semesters, with the implementation of RAG representing the principal instructional difference between cohorts. This design offers a realistic evaluation of RAG-enhanced AI feedback under authentic instructional conditions while minimising changes to the broader learning environment. Maintaining these instructional conditions across adjacent semesters increases confidence that observed differences are more plausibly associated with the implementation of RAG than with changes in course design or instructional practice.

The underlying large language model remained the same across cohorts. The principal technical difference between semesters was the implementation of RAG, which grounded generated feedback in course-specific instructional resources, prior student exemplars, instructor materials, and other curated knowledge sources. No other substantive changes were made to the AI feedback workflow during the study period.

The primary strength of this design lies in its ecological validity. Rather than evaluating RAG under artificial laboratory conditions, the study examined its implementation within a live graduate course where students engaged in authentic writing, feedback, revision, and assessment activities. Although this design sacrifices some experimental control, it provides stronger evidence regarding how RAG-enhanced AI feedback functions in realistic educational settings.

Participants

We compared responses from two successive course cohorts to examine differences in learners' perceptions between the non-RAG and RAG conditions. The Fall 2023 cohort (n = 96) completed the survey under the non-RAG condition, whereas the Spring 2024 cohort (n = 86) completed the same survey under the RAG condition. Both cohorts were taught by the same instructional team and completed the same course using identical assignments, assessment rubrics, essay-writing workflows, and post-course perception surveys. The primary instructional difference between semesters was the implementation of RAG within the AI feedback system. Demographic characteristics were broadly comparable across cohorts:

- Gender: Fall cohort: 72% female; Spring cohort: 68.6% female
- Geography: Fall cohort: 84% North American; Spring cohort: 84.88% North American
- Educational Level: Fall cohort: 2.08% certificate; 59.38% Master's; 38.54% Doctorate; Spring cohort: 1.16% other; 1.16% Non-degree; 4.65% Bachelor's; 56.98% Master's; 33.72% Doctorate

Participants' ages ranged from 18 to over 55 years. For the educational-level comparison, qualifications were grouped into master's, doctorate, and other qualifications. Pearson chi-square tests, without continuity correction, detected no statistically significant cohort differences in the female/male gender distribution, $\chi^2(1, N = 179) = 0.13$, $p = .723$; North American versus non-North-American location, $\chi^2(1, N = 181) = 0.08$, $p = .776$; or educational-level distribution, $\chi^2(2, N = 182) = 4.64$, $p = .098$. The gender comparison excluded one non-binary response and two responses indicating non-disclosure; the geographic comparison excluded one non-disclosed location. These findings do not establish cohort equivalence or eliminate the possibility of unmeasured cohort, historical, instructional, or contextual differences between terms.

Intervention Protocol

The intervention involved integrating Retrieval-Augmented Generation (RAG) into CGScholar's AI feedback system during Spring 2024. In this study, RAG refers to a process in which the system retrieves relevant external materials before generating feedback so that its responses are grounded in course-related and semantically similar sources rather than relying only on the base language model. Embedding-based vector matching supported this retrieval process. Put simply, texts were converted into numerical representations of semantic meaning so the system could identify passages that were conceptually related to a student submission even when they did not share the same keywords.

The RAG system drew on a vector database containing more than 35 million tokens derived from prior student submissions, instructor-authored materials, and a curated set of 117,705 peer-reviewed documents. Source texts were segmented into chunks of approximately 300–500 words and converted into 1,536-dimensional embeddings. During inference, the system used cosine similarity to identify the most relevant records and retrieved the top seven matches for inclusion in the prompt context. These settings were selected as practical design decisions for this instructional environment rather than as universally optimal retrieval parameters. The goal was to strike a workable balance between contextual specificity, semantic coherence, and prompt economy: chunks large enough to preserve local meaning, but not so large that retrieval became diffuse; and enough retrieved material to ground the response, but not so much that competing context crowded the prompt.

Data Collection Instruments

Quantitative data were collected through structured post-course surveys administered via Qualtrics at the end of each semester. The survey asked students to rate their perceptions of AI-generated feedback quality, usefulness, and actionability using single 5-point Likert-type items (1-5; higher scores indicate more positive perceptions). The focal items asked respondents to rate (1) the overall quality of the AI review created through CGScholar (1 = Very poor to 5 = Excellent), (2) the usefulness of the AI review in improving their work based on the course rubric (1 = Not at all useful to 5 = Excellently useful), and (3) the actionability of the feedback, that is, how easy or difficult it was to implement the AI review's suggestions during revision (1 = Very difficult to 5 = Very easy).

Multi-item measures are often preferable when a construct is broad or highly layered, but single-item measures can perform adequately when the construct is narrow, concrete, and relatively unambiguous (Allen et al., 2022; Wanous et al., 1997). They are also commonly used in educational technology research to capture focused perceptual judgments such as perceived usefulness and satisfaction (Chen & Liu, 2020). In this study, the three items were intended as concise indicators of students' end-of-course judgments in each area rather than as full operationalisations of multidimensional constructs. The findings should therefore be read as evidence about overall perceptions, not as exhaustive measures of the full dimensionality of feedback quality, usefulness, or actionability. Additional survey items captured demographic information, including age, gender, educational level pursued, and geographic location.

Data Analysis Framework

The analytical strategy was designed to examine whether learners' perceptions differed between cohorts while accounting for measured demographic characteristics and remaining consistent with the quasi-experimental design. The analysis proceeded in several steps. Cronbach's alpha was calculated across the three perception items to provide a descriptive indication of internal consistency within each cohort (.79 for Fall 2023 and .84 for Spring 2024). We report alpha here only as an indicator of response coherence across the broader perception battery, not as evidence that the three items form a single scale. Because quality, usefulness, and actionability were conceptualised and analysed as related but distinct dimensions, alpha is included as a descriptive survey-level statistic rather than as a justification for aggregation.

Second, because the study compared two independent cohorts, Welch independent-samples t-tests were used to compare Fall 2023 and Spring 2024 on perceived feedback quality, usefulness, and actionability. A Bonferroni-adjusted alpha of .017 was applied to account for multiple comparisons. Cohen's d was calculated to estimate practical significance, using the conventional thresholds of .20 for small effects, .50 for medium effects, and .80 for large effects.

Third, regression analyses were conducted for each outcome variable to examine whether between-cohort differences remained after accounting for age, gender, geographic location, and educational level. These models were used to estimate association patterns in the observed data rather than to establish causal effects of the RAG implementation. Assumption checks indicated no major concerns regarding multicollinearity (VIF < 2.5), residual normality (Shapiro-Wilk $p > .05$), or homoscedasticity (Breusch-Pagan $p > .05$). A G*Power analysis indicated sufficient sensitivity to detect medium effect sizes ($d = .50$) at alpha = .05 with power = .80.

Ethical Considerations

The study adhered to strict ethical standards throughout its design and implementation phases. IRB approval was obtained prior to data collection (#17887). Participants provided informed consent before engaging in study activities, ensuring voluntary participation without coercion or academic repercussions.

Results

Descriptive data in the form of means of ratings were generated and compared to show differences in perception between the Fall 2023 (non-RAG) and Spring 2024 (RAG) groups (Figure 1).

Figure 1. Comparison of Fall 2023 and Spring 2024 Means

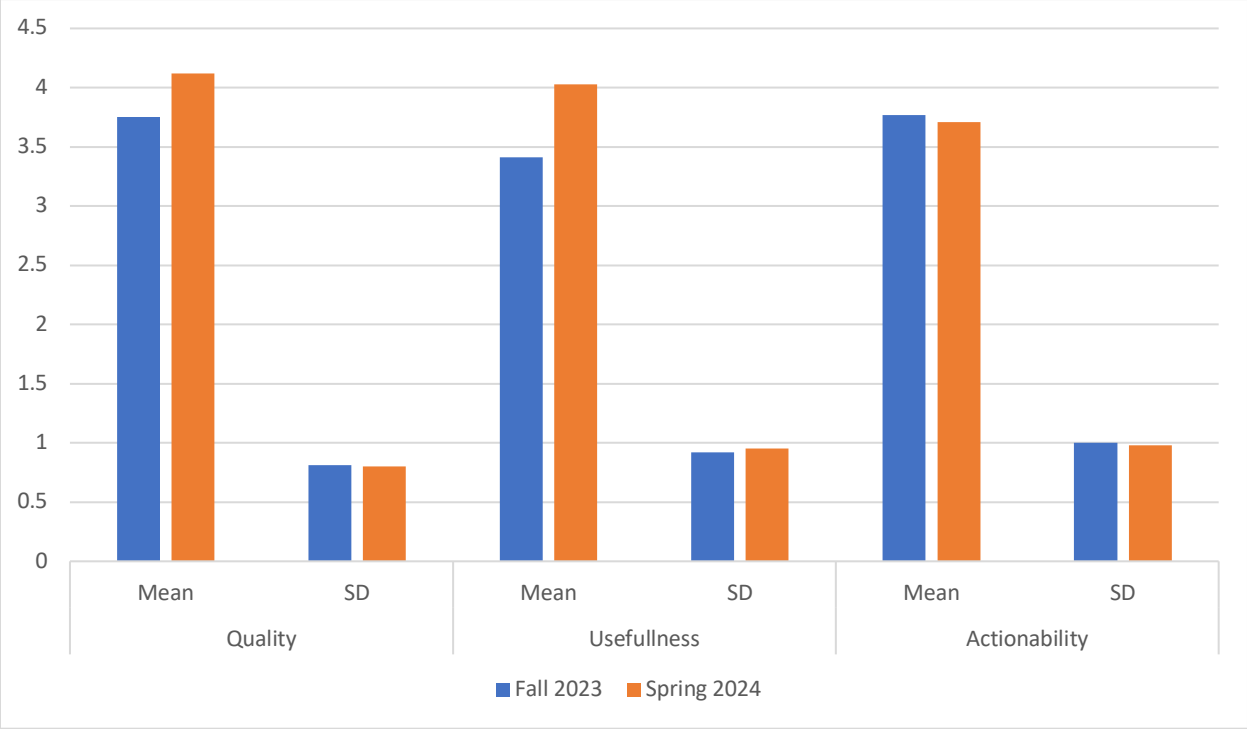


Figure 1 provides a simple visual comparison of mean ratings across cohorts for perceived quality, usefulness, and actionability. The RAG cohort reported higher mean ratings for perceived quality (4.12 vs. 3.75) and usefulness (4.03 vs. 3.41). Perceived actionability showed almost no movement across cohorts and was slightly lower in Spring 2024 (3.71 vs. 3.77). Descriptively, the RAG cohort reported higher ratings of quality and usefulness, whereas actionability ratings were nearly identical across cohorts. Inferential analyses were conducted to determine whether these differences exceeded what might reasonably be expected from sampling variability.

To examine whether those visible differences were statistically meaningful, a series of inferential analyses was conducted. We begin with perceived quality (Table 1).

Group	*N*	*Mean*	*SD*	*t*	*df*	*p*	*d*
Fall ’23	96	3.75	0.81				
Spring ’24	86	4.12	0.80	3.06	178	.003	0.45

Table 1. Quality of feedback

The t-test revealed a statistically significant difference in quality ratings between the two cohorts, t (178) = 3.06, p = .003. Students in Spring 2024 (M = 4.12, SD = 0.80) rated the feedback as higher in quality than students in Fall 2023 (M = 3.75, SD = 0.81). The effect size (d = 0.45) indicates a small-to-moderate difference, suggesting that the RAG cohort experienced the feedback as more clearly expressed, specific, and understandable.

We then tested for differences in perceived usefulness (Table 2).

Group	*N*	*Mean*	*SD*	*t*	*df*	*p*	*d*
Fall ’23	96	3.41	0.92				
				4.51	177	< .001	0.67
Spring ’24	86	4.03	0.95				

Table 2. Usefulness of feedback

The results showed a statistically significant difference in perceived usefulness between the two groups, t (177) = 4.51, p < .001. Students in Spring 2024 (M = 4.03, SD = 0.95) rated the tool as more useful than students in Fall 2023 (M = 3.41, SD = 0.92). The effect size (d = 0.67) falls in the medium-to-large range, suggesting that the addition of RAG was associated with a moderate between-cohort difference in perceived usefulness.

Finally, we examined perceived actionability (Table 3).

Group	*N*	*Mean*	*SD*	*t*	*df*	*p*	*d*
Fall ’23	96	3.77	1.00				
				−0.42	179	.68	−0.06
Spring ’24	86	3.71	.98				

Table 3. Actionability of feedback

Here, the pattern was different. There was no statistically significant difference in perceptions of actionability between the two cohorts, t (179) = −0.42, p = .68. The mean rating in Fall 2023 (M = 3.77, SD = 1.00) was only slightly higher than in Spring 2024 (M = 3.71, SD = 0.98), and the effect size (d = −0.06) was negligible. The analysis provided no evidence of a between-cohort difference in students’ perceptions of actionability.

We also estimated exploratory multiple linear regression models to examine whether cohort differences in students’ perceptions remained after adjustment for measured demographic characteristics. Separate models were estimated for perceived feedback quality, usefulness, and actionability. Each model included cohort, age category, gender, geographic location, and educational level as predictors. The models were used to describe adjusted associations in the observed data; they do not isolate the causal effect of RAG or rule out unmeasured differences between cohorts.

Cohort was coded 0 for Fall 2023 and 1 for Spring 2024. Gender was coded 0 for female and 1 for male. Location was coded 0 for North America and 1 for outside North America. Age was entered as an ordered five-category variable, and educational level was coded from 0 to 4:

other/non-credit, certificate, bachelor's, master's, and doctorate. Tables 4–6 report unstandardized regression coefficients (B), standard errors, t-statistics, and p-values.

	Coefficients					
Predictors	*B*	*SE*	R^2	*F-change*	*t-value*	*p-value*
(Constant)	2.80	0.29	0.11	4.32	6.17	< .001
Intervention	0.41	0.13			3.01	.003
Age	0.01	0.06			0.17	0.86
Gender	−0.02	0.12			−0.18	0.85
Location	−0.03	0.08			0.40	0.69
Education	0.28	0.09			3.18	0.002

Table 4. Regression analysis of *quality*

The overall model predicting perceived feedback quality was statistically significant, R^2 = .11, adjusted R^2 = .09, F (5, 176) = 4.32, p < .001. Together, the five predictors accounted for approximately 11% of the observed variation in quality ratings.

After adjustment for the measured demographic characteristics, membership in the Spring 2024 cohort was associated with a 0.41-point higher perceived-quality rating than membership in the Fall 2023 cohort, B = 0.41, SE = 0.13, t = 3.01, p = .003. Educational level was also positively associated with quality ratings, B = 0.28, SE = 0.09, t = 3.18, p = .002. Age category, gender, and location were not statistically associated with quality ratings in this model.

Thus, the observed cohort difference in quality ratings remained after adjustment for the measured demographic variables. This adjusted association should not be interpreted as evidence that RAG caused the difference, because cohort membership remained confounded with academic terms and other unmeasured contextual differences.

We then estimated the same type of model for perceived usefulness (Table 5).

	Coefficients					
Predictors	*B*	*SE*	R^2	*F-change*	*t-value*	*p-value*
(Constant)	2.27	0.34	.17	7.03	3.77	< .001
Intervention	0.61	0.16			3.90	< .001
Age	0.07	0.07			0.96	0.34
Gender	0.12	0.14			0.85	0.40
Location	−0.09	0.10			−0.97	0.33
Education	0.30	0.10			2.91	0.004

Table 5. Regression analysis of *usefulness*

The overall model predicting perceived feedback usefulness was statistically significant, R^2 = .17, adjusted R^2 = .15, F (5, 176) = 7.03, p < .001. The predictors collectively accounted for approximately 17% of the observed variation in usefulness ratings.

Membership in the Spring 2024 cohort was associated with a 0.61-point higher perceived-usefulness rating after adjustment for the measured demographic characteristics, $B = 0.61$, $SE = 0.16$, $t = 3.90$, $p < .001$. Educational level was also positively associated with usefulness ratings, $B = 0.30$, $SE = 0.10$, $t = 2.91$, $p = .004$. Age category, gender, and location were not statistically associated with usefulness ratings.

The adjusted cohort association therefore remained evident for usefulness. As with the quality model, this finding should be interpreted as an association between cohort membership and perceived usefulness after adjustment for the measured covariates, not as a causal estimate of the RAG implementation.

A final regression model was then estimated for perceived actionability (Table 6).

	Coefficients					
Predictors	*B*	*SE*	R^2	*F-change*	*t-value*	*p-value*
(Constant)	3.10	0.35			5.92	.008
Intervention	0.16	0.16			0.99	.32
Age	−0.18	0.07	.08	3.22	−2.36	.02
Gender	0.12	0.14			0.87	.38
Location	−0.24	0.10			−2.32	.02
Education	0.30	0.11			2.74	.01

Table 6. Regression analysis of *actionability*

The overall model predicting perceived actionability was statistically significant, $R^2 = .08$, adjusted $R^2 = .05$, F (5, 176) = 3.22, $p = .008$. The predictors collectively accounted for approximately 8% of the observed variation in actionability ratings.

Cohort membership was not statistically associated with actionability ratings after adjustment for the measured demographic characteristics, $B = 0.16$, $SE = 0.16$, $t = 0.99$, $p = .32$. Higher age-category scores were negatively associated with actionability ratings, $B = -0.18$, $SE = 0.07$, $t = −2.36$, $p = .02$, as was the non-North-American location category, $B = −0.24$, $SE = 0.10$, $t = −2.32$, $p = .02$. Educational level was positively associated with actionability ratings, $B = 0.30$, $SE = 0.11$, $t = 2.74$, $p = .01$. Gender was not statistically associated with actionability.

These model-based associations should be interpreted cautiously. In particular, the cohort coefficient did not provide evidence of an adjusted difference in perceived actionability, and the demographic associations are observational rather than causal.

Discussion

This study examined whether the introduction of RAG-enhanced AI feedback was associated with graduate students' perceptions of feedback quality, usefulness, and actionability within an authentic course-based cyber-social learning environment deliberately shaped by learning-by-design as an educational philosophy. That matters because we were not trying to show everything feedback does. We were trying to understand one part of a larger picture: what appears to change when a technical intervention is introduced into an already pedagogically intentional environment. The findings revealed an uneven pattern: students in the RAG cohort reported higher perceived

quality and usefulness, whereas perceived actionability did not differ between cohorts. This pattern suggests that improving the grounding and contextual relevance of AI-generated feedback may influence how learners evaluate the feedback without necessarily changing their ability to implement it.

Perceived Feedback Quality (RQ1)

RQ1 examined whether perceived feedback quality differed between the non-RAG and RAG cohorts. The effect size, Cohen's d = 0.45, indicates a small-to-moderate difference between the cohorts. Put plainly, students in the RAG group experienced the AI reviews as more clearly expressed, more specific, and easier to understand.

From a cyber-social learning perspective, that improvement makes sense. It aligns most closely with the transpositional and reflexive precepts. Because the RAG system grounded generated feedback in human-authored artefacts, including prior student work, instructor materials, and peer-reviewed sources, students in the RAG cohort may have experienced the commentary as more contextually relevant to the task at hand. A grounded system is more likely to produce feedback that sounds situated rather than generic, and that difference may matter a great deal at the level of perception. Although hallucination frequency was not measured, one possibility is that grounding reduced irrelevant or weakly contextualised comments, which may have contributed to the higher quality ratings. This interpretation should be examined directly in future work. The positive association between educational level and perceived quality adds another layer to this interpretation. Students with more advanced academic experience may be better positioned to notice and value these transpositional gains, which fits with research linking feedback literacy to the ability to interpret and evaluate feedback well (Henderson et al., 2019; Winstone & Carless, 2019). It may also be that students who have moved through more programmes of study, and therefore through more cycles of giving and receiving feedback, are more sensitive to shifts in quality when they occur.

Perceived Feedback Usefulness (RQ2)

RQ2 examined how a RAG intervention affects perceptions of the usefulness of AI-generated feedback. Here the pattern was even a little stronger. The effect size (Cohen's d = 0.67) indicates a medium-to-large difference between cohorts, and the regression model showed that both the RAG cohort indicator and educational level were positively associated with perceived usefulness.

Within the cyber-social learning framework, usefulness can be understood in relation to how well feedback supports a learner's participation in the reflexive cycle and how successfully transposed knowledge resources are made meaningful for human interpretation. Grounding AI feedback in course-specific and discipline-relevant materials may have made the comments feel more aligned with students' actual goals, which would help explain the stronger usefulness ratings. The fact that more advanced students also reported greater usefulness echoes prior work linking higher educational levels to more positive views of AI and to stronger feedback literacy (Mukhamedkarimova & Umurkulova, 2025). For these learners, RAG-enhanced feedback may have offered richer cues for self-regulation and revision planning. At the same time, it is important not to overstate what RAG is doing. RAG-enhanced feedback is not dialogic in itself, if dialogue is understood as reciprocal human meaning-making. Its contribution is more indirect. By producing commentary that is better grounded and more locally relevant, it may provide better

material for later interpretation, questioning, and discussion among students, peers, and instructors. In that sense, RAG may support a dialogic feedback process without replacing the human interaction through which feedback is actually negotiated and enacted.

Perceived Feedback Actionability (RQ3)

RQ3 examined whether perceived actionability differed between the RAG and non-RAG cohorts. In contrast to quality and usefulness, there was no statistically significant difference in actionability ratings between cohorts, and the effect size was negligible. The regression model tells the same story. Although the overall model was statistically significant, the intervention variable itself was not a significant predictor once age, gender, location, and education were controlled.

In the adjusted model, age and non-North American location were negatively associated with actionability, whereas educational level was positively associated with it. These secondary findings should be interpreted cautiously because they were based on observational covariates, involved multiple statistical tests, and were not the study's primary hypotheses. Nevertheless, they identify potentially important questions about whether learners' prior experience, institutional context, or familiarity with dominant academic conventions influences how readily AI-generated feedback can be translated into revision.

Cyber-social learning theory helps explain why RAG did not improve actionability even though it appeared to strengthen quality and usefulness. Whereas quality and usefulness are shaped most directly by the transpositional and reflexive dimensions of the learning ecology, actionability seems to depend more heavily on the collaborative precept, that is, on the social and dialogic work through which learners interpret feedback, test its meaning, and decide what to do next. A RAG-enhanced system can make feedback more factually grounded and more contextually relevant, but it cannot by itself provide the relational, affective, and culturally responsive support that often makes feedback actionable in practice (Al-Obaydi & Pikhart, 2025; Fong & Schallert, 2023). The fact that age, education, and location mattered here is therefore theoretically meaningful. The exploratory associations with age, education, and location are consistent with the possibility that actionability depends partly on learner experience and contextual fit, although further research is needed to test these explanations. If actionability is the goal, pedagogical scaffolding and attention to cultural and regional variation still matter very much.

Implications and Future Research

These findings provide preliminary empirical support for viewing RAG as one component of a hybrid feedback system by showing more precisely where students appeared to experience benefits from RAG-enhanced AI feedback and where they did not. From a cyber-social learning perspective, the gains in perceived quality and usefulness suggest that strengthening the transpositional and reflexive dimensions of the learning ecology can improve students' evaluative judgments of AI-generated feedback in meaningful ways. At the same time, the absence of change in perceived actionability is consistent with research suggesting that relevant feedback does not automatically translate into revision; learners may also require opportunities to clarify feedback, negotiate its meaning, develop action plans, and monitor their subsequent work through collaborative and dialogic processes (Carless & Boud, 2018; Er et al., 2021; Nicol, 2010). This interpretation also aligns with emerging research in design education, where RAG-supported AI

review has been positioned as a complement to reflective and instructor-supported feedback processes rather than as a replacement for them (Hughes et al., 2024).

For educators and institutions, the practical implication is that RAG should be treated as one part of a broader feedback architecture rather than as a standalone solution. In courses where students already have relatively strong feedback literacy, such as graduate and doctoral programmes, RAG appears associated with more positive perceptions of feedback precision and relevance. But that does not mean technical enhancement is enough. The negative association between age and perceived actionability, for example, suggests that some learners may need additional support in interpreting and implementing AI-generated suggestions. The exploratory age association indicates a need to investigate whether support requirements vary across learner groups; it does not by itself establish that age-specific support is necessary. Put another way, better grounding does not automatically close gaps in feedback use across different learner groups.

These findings and the cyber-social framework suggest several design principles for future implementation and evaluation:

- Design RAG as a reflexive node in a hybrid feedback system, pairing AI reviews with peer and instructor feedback cycles rather than replacing existing human feedback. This can help ensure that students receive both contextually grounded commentary from the AI and relational, dialogic support from humans.
- Embed explicit feedback literacy activities into curricula, especially for less advanced learners, to support interpretation and use of AI-generated comments. Examples include structured reflection prompts asking students to compare AI, peer, and instructor feedback, or guided workshops where students practise revising work in response to multi-source feedback.
- Integrate actionability checks into assessment design, such as requiring students to document how they acted on AI feedback in subsequent drafts, thereby making the implementation of feedback a visible and assessable part of coursework.

For developers and learning technology teams, the results point to design priorities that align RAG systems more closely with cyber-social learning principles:

- Strengthen transpositional transparency by providing interface elements that show which human-authored sources informed particular feedback statements (e.g. course readings, exemplar assignments), thereby helping students understand and evaluate the provenance of AI suggestions.
- Explore learner-controlled options that calibrate the complexity, specificity, and tone of AI feedback to learners' educational level and previous interactions, recognising that advanced students may seek more nuanced critique while less advanced students may require more scaffolded, step-by-step guidance rather than assuming that educational level alone determines feedback needs.
- Explore features that support collaborative interpretation, such as tools that allow students to annotate AI feedback, share it with peers, and bring it into tutorials or online discussions, so that actionability emerges from social negotiation rather than from AI output alone.

Future studies should directly evaluate retrieval quality, source relevance, and hallucination frequency so that changes in student perceptions can be linked to verified changes in system

performance. Research should also examine revision behaviour and document-level changes to determine whether perceived usefulness and actionability correspond to actual feedback uptake, triangulating this perceptual data with behavioural and learning outcomes across more diverse institutional and cultural contexts. Longitudinal and mixed-methods studies would be especially valuable here, as they could show not only whether students like or trust AI feedback at a given moment, but how their relationship with it changes over time. That kind of work would help clarify how cyber-social learning environments can be intentionally designed to support different learners, rather than assuming the same feedback system will work equally well for everyone.

Limitations

This study has several limitations that shape how the findings should be interpreted. First, the non-equivalent cohort comparison design supports only cautious causal inference, because differences between semesters may reflect unmeasured cohort or contextual effects as well as the RAG implementation. Second, the study was conducted within a single platform-based graduate learning environment at one institution, which limits transferability to other settings and learner populations. Third, variables such as prior AI experience, feedback literacy, disciplinary background, and attitudes toward automated feedback were not measured and may have differed between cohorts. Finally, the study did not include instructor perceptions, revision behaviour, or learning outcomes. The findings should therefore be understood as evidence about students' perceived experiences of AI-generated feedback rather than as definitive evidence of improved pedagogical effectiveness.

Relatedly, each perception dimension was measured with a single survey item. While single-item measures can be acceptable for narrow and relatively unambiguous constructs (Wanous et al., 1997), they do not allow for assessment of within-construct reliability and they are less able to capture the breadth of multidimensional constructs such as feedback quality. Future research would benefit from multi-item validated scales for each perception dimension so that psychometric evaluation can be more robust and the underlying measurement model can be examined more directly.

Conclusions

This study examined whether the implementation of RAG was associated with differences in graduate students' perceptions of AI-generated feedback. The findings suggest a clear but uneven pattern. Students in the RAG cohort reported higher perceived quality and usefulness, whereas perceived actionability did not differ between cohorts. The higher ratings are consistent with the possibility that grounding made the feedback appear more contextually relevant or credible, although the study did not directly evaluate retrieval quality or generated-feedback accuracy. However, RAG was not associated with improved perceptions of actionability. This uneven pattern reinforces the distinction between improving the informational characteristics of AI feedback and supporting learners in interpreting and implementing it. That matters because it suggests that better information retrieval alone does not solve the harder pedagogical problem of helping students decide what to revise and how to revise it. For researchers, the findings point to the importance of studying complementary approaches that support actionability, including scaffolding, feedback literacy development, and the emotional and relational dimensions of feedback. Future work should also combine technical evaluation, validated perceptual measures,

revision behaviour, and learning outcomes. For practitioners, especially in higher education, the results suggest that RAG may improve how students perceive AI-generated feedback, but that it works best when embedded in a broader human-supported feedback ecology.

Acknowledgements

The authors disclose that they have no actual or perceived conflicts of interest. The authors disclose that they have not received any funding for this manuscript beyond resourcing for academic time at their respective universities. The authors confirm that they have met the ethical standards expected. The authors have used artificial intelligence in this manuscript and meet expected principles. The lead author has used Chat GPT-5.6 Terra through the Perplexity AI platform to proofread and condense writing for clarity and to help with confirming reference lists are accurately reported. Writing has been checked against APA 7 and JUTLP writing and formatting guidelines.

The authors list the following CRediT contributions:
Christopher Hughes: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft. Akash Saini: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – review & editing. Mary Kalantzis: Conceptualization, Writing – review & editing. Bill Cope: Conceptualization, Writing – review & editing

References

Allen, M., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science: A call to action. *European Journal of Psychological Assessment*, *38*(1), 1–2. https://doi.org/10.1027/1015-5759/a000699

Al-Obaydi, L. H., & Pikhart, M. (2025). AI partner versus human partner: Comparing AI-based peer assessment with human-generated peer assessment in examining writing skills. *Language Testing in Asia, 15*, Article 38. https://doi.org/10.1186/s40468-025-00375-8

Arowosegbe, A., Alqahtani, J. S., & Oyelade, T. (2024). Perception of generative AI use in UK higher education. *Frontiers in Education, 9*, Article 1463208. https://doi.org/10.3389/feduc.2024.1463208

Balaguer, A., Benara, V., Cunha, R. L., Filho, R. D., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., & Chandra, R. (2024). *RAG vs fine-tuning: Pipelines, trade-offs, and a case study on agriculture* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2401.08406

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*, Article 43. https://doi.org/10.1186/s41239-023-00411-8

Chen, L.-T., & Liu, L. (2020). Methods to analyze Likert-type data in educational technology research. *Journal of Educational Technology Development and Exchange*, *13*(2), 39–60. https://doi.org/10.18785/jetde.1302.04

Cope, B., & Kalantzis, M. (2015). The things you do to know: An introduction to the pedagogy of multiliteracies. In B. Cope & M. Kalantzis (Eds.), *A pedagogy of multiliteracies: Learning by design* (pp. 1–36). Palgrave Macmillan. https://doi.org/10.1057/9781137539724_1

Cope, B., & Kalantzis, M. (2019). Education 2.0: Artificial intelligence and the end of the test. *Beijing International Review of Education*, *1*(2), 528–543. https://doi.org/10.1163/25902539-00102009

Cope, B., & Kalantzis, M. (2022a). The cybernetics of learning. *Educational Philosophy and Theory*, *54*(14), 2352–2388. https://doi.org/10.1080/00131857.2022.2033213

Cope, B., & Kalantzis, M. (2022b). Artificial intelligence in the long view: From mechanical intelligence to cyber-social systems. *Discover Artificial Intelligence*, *2*(13), 1–18. https://doi.org/10.1007/s44163-022-00029-1

Cope, B., & Kalantzis, M. (2024). On cyber-social learning: A critique of artificial intelligence in education. In D. Kourkoulou, A. O. Tzirides, B. Cope, & M. Kalantzis (Eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines* (pp. 3–34). Springer. https://doi.org/10.1007/978-3-031-64487-0_1

Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*, *53*(12), 1229–1245. https://doi.org/10.1080/00131857.2020.1728732

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

Dakshit, S. (2024). Faculty perspectives on the potential of RAG in computer science higher education. In *The 25th Annual Conference on Information Technology Education (SIGITE '24)* (pp. 19–24). Association for Computing Machinery. https://doi.org/10.1145/3686852.3686864

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment*, *5*(1), 1–35.

Elmessiry, A., & Elmessiry, M. (2024). Navigating the evolution of artificial intelligence: Towards education-specific retrieval augmented generative AI (ES-RAG-AI). In *INTED2024 Proceedings* (pp. 7692–7697). IATED. https://doi.org/10.21125/inted.2024.2052

Er, E., Dimitriadis, Y., & Gašević, D. (2021). A collaborative learning approach to dialogic peer feedback: A theoretical framework. *Assessment & Evaluation in Higher Education*, *46*(4), 586–600. https://doi.org/10.1080/02602938.2020.1786497

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). *A survey on RAG meeting LLMs: Towards retrieval-augmented large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2405.06211

Fong, C. J., & Schallert, D. L. (2023). “Feedback to the future”: Advancing motivational and emotional perspectives in feedback research. *Educational Psychologist*, *58*(3), 146–161. https://doi.org/10.1080/00461520.2022.2134135

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.10997

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). *A survey on large language models: Applications, challenges, limitations, and practical usage* [Preprint]. TechRxiv. https://doi.org/10.36227/techrxiv.23589741.v1

Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, *90*, Article 104512. https://doi.org/10.1016/j.ebiom.2023.104512

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

Haughney, K., Wakeman, S., & Hart, L. (2020). Quality of feedback in higher education: A review of literature. *Education Sciences*, *10*(3), Article 60. https://doi.org/10.3390/educsci10030060

Henderson, M., Phillips, M., Ryan, T., Boud, D., Dawson, P., Molloy, E., & Mahoney, P. (2019). Conditions that enable effective feedback. *Higher Education Research & Development*, *38*(7), 1401–1416. https://doi.org/10.1080/07294360.2019.1657807

Hughes, C., Tzirides, A. O., & Saini, A. K. (2024). A quantitative investigation of graduate student perceptions of human-generated and AI-generated reviews in a cyber-social learning platform. In D. Kourkoulou, A. O. Tzirides, B. Cope, & M. Kalantzis (Eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines* (pp. 213–234). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64487-0_10

Jauhiainen, J. S., & Garagorry Guerra, A. (2024). Generative AI in education: ChatGPT-4 in evaluating students’ written responses. *Innovations in Education and Teaching International*, *62*(4), 1377–1394. https://doi.org/10.1080/14703297.2024.2422337

Jia, K., Wang, P., Li, Y., Chen, Z., Jiang, X., Lin, C.-L., & Chin, T. (2022). Research landscape of artificial intelligence and e-learning: A bibliometric research. *Frontiers in Psychology*, *13*, Article 795039. https://doi.org/10.3389/fpsyg.2022.795039

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Klyshbekova, M., & Forsberg, A. (2026). ‘Learning to use AI to help us but not to do it for us’: Understanding UK undergraduate students’ use of GenAI. *Cogent Education, 13*(1), Article 2650254. https://doi.org/10.1080/2331186X.2026.2650254

Koohang, A., Sargent, C. S., & Svanadze, S. (2024). Students’ perceptions of benefits and opportunities of artificial intelligence (AI). *Issues in Information Systems, 25*(2), 438–450. https://doi.org/10.48009/2_iis_2024_134

Miladi, F., Psyché, V., & Lemire, D. (2024). Leveraging GPT-4 for accuracy in education: A comparative study on retrieval-augmented generation in MOOCs. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (Communications in Computer and Information Science, Vol. 2150, pp. 427–434). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64315-6_40

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mukhamedkarimova, D. F., & Umurkulova, M. M. (2025). Factors contributing to higher education students' acceptance of artificial intelligence: A systematic review. *European Journal of Educational Research, 14*(4), 1373–1388. https://doi.org/10.12973/eu-jer.14.4.1373

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2024). AI or human? Evaluating student feedback perceptions in higher education. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruipérez Valiente (Eds.), Technology enhanced learning for inclusive and equitable quality education (pp. 284–298). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72315-5_20

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education, 35*(5), 501–517. https://doi.org/10.1080/02602931003786559

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

O'Dell, M., & DeOrio, A. (2025, June). *Specialist vs. generalist generative AI for student learning*. In *2025 ASEE Annual Conference & Exposition*. American Society for Engineering Education. https://doi.org/10.18260/1-2--57666

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review, 35*, Article 100416. https://doi.org/10.1016/j.edurev.2021.100416

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, *50*(1), 128–138. https://doi.org/10.1111/bjet.12592

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE*, *19*(5), Article e0304013. https://doi.org/10.1371/journal.pone.0304013

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology, 108*, 1098–1120. https://doi.org/10.1037/edu0000103

Rangan, K., & Yin, Y. (2024). A fine-tuning enhanced RAG system with quantized influence measure as AI judge. *Scientific Reports*, *14*, Article 27446. https://doi.org/10.1038/s41598-024-79110-x

Rashid, S. F., Duong-Trung, N., & Pinkwart, N. (2024). *Generative AI in education: Technical foundations, applications, and challenges*. IntechOpen. https://doi.org/10.5772/intechopen.1005402

Rietsche, R., & Söllner, M. (2019). Insights into using IT-based peer feedback to practice the students providing feedback skill. *Hawaii International Conference on System Sciences 2019 (HICSS-52)*. https://aisel.aisnet.org/hicss-52/cl/teaching_and_learning_technologies/8

Saini, A. K., Tzirides, A. O., Cope, B., Kalantzis, M., & Searsmith, D. (2025). Generative AI and the next generation of feedback systems: A hybrid approach for higher education. In *Proceedings of the 19th International Conference of the Learning Sciences (ICLS 2025)* (*pp. 2120–2124*). International Society of the Learning Sciences. https://doi.org/10.22318/icls2025.631446

Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner–instructor interaction in online learning. *International Journal of Educational Technology in Higher Education*, *18*, Article 54. https://doi.org/10.1186/s41239-021-00292-9

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Taylor & Francis.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Stöhr, C., Ou, A. W., & Malmström, H. (2024). Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. *Computers and Education: Artificial Intelligence, 7*, Article 100259. https://doi.org/10.1016/j.caeai.2024.100259

Thüs, D., Malone, S., & Brünken, R. (2024). Exploring generative AI in higher education: A RAG system to enhance student engagement with scientific literature. *Frontiers in Psychology*, *15*, Article 1474892. https://doi.org/10.3389/fpsyg.2024.1474892

Tzirides, A. O., Saini, A. K., Cope, B., Kalantzis, M., & Searsmith, D. (2023). Cyber-social research: Emerging paradigms for interventionist education research in the postdigital era. In P. Jandrić, A. MacKenzie, & J. Knox (Eds.), *Constructing postdigital research: Method and emancipation* (pp. 85–102). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35411-3_5

Tzirides, A. O., Saini, A. K., Zapata, G., Searsmith, D., Cope, B., Kalantzis, M., Castro, V., Kourkoulou, D., Jones, J. W., da Silva, R. A., Whiting, J., & Kastania, N. P. (2024). Generative AI application in higher education student work. In D. Kourkoulou, A. O. Tzirides, B. Cope, & M. Kalantzis (Eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines* (pp. 287–301). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64487-0_13

Van der Kleij, F. M., & Lipnevich, A. A. (2021). Student perceptions of assessment feedback: A critical scoping review and call for research. *Educational Assessment, Evaluation and Accountability*, *33*, 345–373. https://doi.org/10.1007/s11092-020-09331-x

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Wang, Z., & Ormerod, C. (2024). *Generative language models with retrieval augmented generation for automated short answer scoring* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2408.03811

Winstone, N. E., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge.

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*, Article 101826. https://doi.org/10.1016/j.cedpsych.2019.101826

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., Hessert, W. T., Williams, M. E., & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General, 143*(2), 804–824. https://doi.org/10.1037/a0033906

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., & Shi, S. (2023). *Siren's song in the AI ocean: A survey on hallucination in large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.01219

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90–102. https://doi.org/10.1016/j.asw.2018.02.004