JOURNAL OF LEARNING ANALYTICS

Volume 12(1), 129–151. https://doi.org/10.18608/jla.2025.8601

AI-Assisted Co-Creation: Bridging Skill Gaps in Student-Generated Content

Stanislav Pozdniakov[1]*, Jonathan Brazil[2], Mehrnoush Mohammadi[3], Mollie Dollinger[4], Shazia Sadiq[5], Hassan Khosravi[6]

Abstract

Engaging students in creating high-quality novel content, such as educational resources, promotes deep and higher-order learning. However, students often lack the necessary training or knowledge to produce such content. To address this gap, this paper explores the potential of incorporating generative AI (GenAI) to review students' work and provide them with real-time feedback and assistance during content creation. Specifically, we use RiPPLE, which enables students to create bite-size learning resources and incorporates instant GenAI feedback, highlighting strengths and suggesting improvements to enhance quality. The AI reviews the resource and provides feedback encompassing three main components: a summary of the resource, a list of strengths, and suggestions for improvement. We evaluate this approach by analyzing log data from 1063 student-created multiple-choice questions (MCQs) and the corresponding AI feedback. This analysis aims to understand the depth, scope, and tone of the feedback provided by the AI, as well as the way students engage with and utilize this feedback in their content creation process. Additionally, we examined the perceived helpfulness of the GenAI feedback analyzed via 3324 student ratings and thematically analyzed 601 comments they provided about the feedback. Our findings demonstrate the potential value of AI-generated feedback for students when integrated into pedagogical design. Our analysis suggests that not only can AI-generated feedback provide students with a breadth of feedback to improve their writing and/or discipline-specific content knowledge, but also it is largely well received by students for both its clarity and its positive tone. Despite challenges in ensuring the accuracy of AI-generated feedback, this study shows how this feedback can enable students to make actionable changes in their academic performance.

Notes for Practice

- Co-authoring educational resources presents a valuable opportunity for instructors to upskill students while simultaneously building a growing repository of materials benefiting both instructors and students.
- One of the impediments for instructors to engage students as co-authors is that students often lack the necessary expertise and confidence to develop high-quality materials independently.
- Instructors could benefit from equipping students with AI assistance by enabling students to receive immediate suggestions and align their work with educational best practices. AI feedback can provide structured guidance to students during the resource creation process, focusing on three critical areas: writing clarity, question design principles, and disciplinary accuracy.
- The approach presented in this paper suggests that instructors thoroughly experiment with prompts, as well as suggesting the need to assess the quality of AI assistance prior to making it accessible to students. While students generally perceive AI assistance as valuable for their co-creation activities, the effectiveness of such systems is currently limited by inconsistencies in feedback quality and accuracy, resulting in students using AI assistance for relatively minor revisions in their work.
- Learning analytics designers planning to implement AI assistance could implement separate prompts and interfaces for ideation and refinement phases, introducing structured reflection prompts to help students critically evaluate AI feedback.

Keywords

Co-creation, feedback, generative AI, learning analytics.

Submitted: 08/08/2024 — **Accepted:** 14/02/2025 — **Published:** 19/03/2025

[1] **Corresponding author Email: s.pozdniakov@uq.edu.au, School of Electrical Engineering and Computer Science, The University of*

Queensland, St Lucia, Queensland, 4072, Australia. ORCID iD: https://orcid.org/0000-0003-4451-9181
[2] Email: j.brazil@uq.edu.au, Institute for Teaching and Learning Innovation, The University of Queensland, St Lucia, Queensland, 4072, Australia. ORCID iD: https://orcid.org/0000-0002-6669-2076
[3] Email: m.mohammadibolbanabad@uq.edu.au, School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, Queensland, 4072, Australia. ORCID iD: https://orcid.org/0000-0001-9596-7414
[4] Email: mollie.dollinger@curtin.edu.au, Faculty of Health Sciences, Curtin University, Western Australia 6845, Australia. ORCID iD: https://orcid.org/0000-0003-1105-9051
[5] Email: shazia@eecs.uq.edu.au, School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, Queensland, 4072, Australia. ORCID iD: https://orcid.org/0000-0001-6739-4145
[6] Email: h.khosravi@uq.edu.au, School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, Queensland, 4072, Australia. ORCID iD: https://orcid.org/0000-0001-8664-6117

1. Introduction

In the dynamic landscape of modern education, traditional teaching paradigms are rapidly evolving to meet the needs of an increasingly interconnected and digitally fluent generation. Central to this evolution is the emergence of new pedagogical approaches that prioritize student engagement, empowerment, and creativity (Buckingham Shum et al., 2019; Dollinger & Lodge, 2018). One recent approach that epitomizes this paradigm shift is the notion of co-creation (Dollinger & Lodge, 2018), which refers to a collaborative process where students actively participate with educators, and often with each other, to create the content and structure of their learning experiences. A popular form of co-creation involves engaging students as active co-creators of knowledge, creating and sharing learning resources with one another (Kim, 2015). Such activities have become increasingly popular and offer two interconnected benefits. First, creating learning content demands deep engagement with course material, promoting high-level cognitive skills as outlined in Bloom's taxonomy (Hilton et al., 2022). Second, leveraging the creative power of many students can result in the rapid and cost-effective creation of large repositories of learning resources that can, in turn, be used for practice and to support personalized learning experiences (Singh et al., 2022). The co-creation of educational resources offers many advantages, yet several significant concerns exist that can impede its effectiveness and broader adoption. These concerns fall into three main categories, and for all three impediments (described in the next paragraph), students' ability to write and communicate their ideas clearly is central.

Many students lack the higher-order cognitive skills necessary for co-creating high-quality educational resources, resulting in materials that lack critical insight and depth (Doyle et al., 2019; Darvishi et al., 2021; Khosravi et al., 2023). In addition, students often lack pedagogical expertise and experience, leading to ineffective multiple-choice questions (MCQs) that fail to assess knowledge or encourage critical thinking (Moore et al., 2021; Khosravi et al., 2023). Furthermore, insufficient disciplinary knowledge can cause the content to misrepresent facts, affecting the credibility and educational value of the resources created (Qian & Lehman, 2017), for which greater scaffolding and feedback are required (Doyle et al., 2019). Some of these concerns can be mitigated through close supervision by educators or by providing targeted workshops and training sessions (Bovill et al., 2011; Lubicz-Nawrocka & Bovill, 2023). However, implementing these solutions becomes challenging in large-scale educational settings where curricula are already tightly packed—a reality in many universities. The logistical constraints of large class sizes and limited instructional time further complicate the effective application of these remedial strategies (Bovill, 2020). One approach to addressing this problem, explored in previous work, involves the use of machine learning (ML) and natural language processing (NLP) techniques to provide automated feedback (Keuning et al., 2019; Cavalcanti et al., 2021). Although traditional AI methods for automated feedback have shown accuracy, they often lacked personalization, offered limited actionable next steps, and typically required meticulous setup and prior model training using course-specific data (Keuning et al., 2019; Cavalcanti et al., 2021). Consequently, previous research also examined how both traditional AI approaches and structured student onboarding could improve students' feedback delivery skills, thereby making them more effective peer reviewers (Darvishi et al., 2022).

Recent advances in generative AI (GenAI) have catalyzed the development of language models capable of generating novel texts and other artifacts from user input. These innovations have attracted significant interest in various educational sectors due to their broad applicability (B. Chen et al., 2023; Denny, Gulwani, et al., 2024). Notable areas of focus include content creation (Denny et al., 2023; Hwang et al., 2024; Jury et al., 2024; Leiker et al., 2023; Choi et al., 2024) and providing feedback (Bernius et al., 2022; Nguyen et al., 2023; Han et al., 2023; Gombert et al., 2024; Pozdniakov et al., 2024; Hutt et al., 2024) using large language models (LLMs). As such, GenAI presents a unique opportunity to guide students during the resource co-creation process at scale while preserving their creative agency and helping them acquire all the necessary skills to develop high-quality resources.

Building on these applications, we explore a pedagogically supported approach to co-creation with AI. In this model, students actively engage in creating content, which not only enhances learning but also allows them to receive constructive feedback on their outputs from LLMs. This integrated approach leverages the strengths of AI to both inspire creative educational endeavours and refine the quality of educational materials through expert AI evaluations. Accordingly, this paper details our research on integrating an AI feedback feature into an existing platform named RiPPLE, which has been developed to enable students to create bite-size learning resources. The AI feedback mechanism is designed with pedagogical and prompting frameworks, ensuring that the feedback provided to students adheres to the best pedagogical practices and effectively uses prompting techniques. We assess the AI feedback approach by examining log data from 1063 student-generated MCQs alongside the corresponding AI feedback. Our analysis focuses on evaluating the depth, scope, and tone of the feedback provided by the AI and investigating how students interact with and apply this feedback in their content creation efforts. Furthermore, we analyzed 3324 student evaluations concerning the usefulness of the AI feedback and conducted a thematic analysis of 601 comments provided by the students regarding the feedback. Our evaluation is guided by the following research questions:

RQ1: What are the characteristics of the AI feedback on student-created resources?

RQ2: How did students engage with and use the feedback provided by the AI?

RQ3: How did students perceive the feedback provided by the AI?

The structure of this paper is outlined as follows. Section 2 reviews existing literature related to co-creation and the application of GenAI in educational contexts. Section 3 describes the RiPPLE platform and details our method for integrating AI feedback into it. Section 4 outlines our evaluation approach and methodology, which is designed to address the proposed research questions. Section 5 discusses the results obtained from our analysis. Finally, Section 6 discusses the implications of the findings, reflects on the research questions, and shares insights gained from incorporating GenAI into an educational setting. Section 7 concludes this paper and summarizes the main implications.

2. Background

In the background section, we review the existing literature on two main themes: co-creation in educational contexts and the application of GenAI for content creation and feedback.

2.1 Co-creation

Over the past decade there has been a shift toward involving multiple stakeholders in the design of learning analytics to give a more active voice to those for whom the innovations are developed (Buckingham Shum et al., 2019; Sanders & Stappers, 2008). This has led to the rise of participatory and co-design approaches, where teachers and students actively shape educational technologies and learning analytics (Dollinger & Lodge, 2018; Sarmiento & Wise, 2022) and cooperatively engage in the co-creation of their educational experience (Dollinger & Lodge, 2020; Bovill, 2020). To facilitate co-creation, which typically requires multiple steps and rounds of feedback from educators, various software platforms have been developed to guide students in resource creation and to support their peers and teachers during the review and revisions of the resource (Kim, 2015; Singh et al., 2022). Despite the introduction of these tools to aid students and teachers, research highlights three main challenges.

Deficiency in higher-order cognitive skills: Co-creation of high-quality resources demands strong higher-order cognitive skills, including critical thinking, analysis, synthesis, and evaluation (Doyle et al., 2019; Darvishi et al., 2021; Moore et al., 2021; Khosravi et al., 2023). Many students, however, may not have developed these complex cognitive capabilities to a sufficient level. This deficiency can result in educational materials that lack critical insight and depth, thereby limiting the effectiveness of the co-created resources.

Lack of pedagogical expertise and experience: A significant concern in the co-creation of educational resources is the general lack of pedagogical expertise among students. For instance, MCQs are a type of resource that is often co-created by students. Creating effective MCQs requires an understanding of various design principles, including clear question framing, the creation of plausible distractors, and alignment with learning objectives. Unfortunately, many students have not been exposed to these principles, resulting in MCQs that may not effectively assess knowledge or encourage critical thinking (Moore et al., 2021; Khosravi et al., 2023). Moreover, students may also face unique barriers due to their lack of expertise or experience in pedagogical design. Depending on students' prior educational experiences, and how actively they were encouraged to exhibit agency in their learning experiences, students may also feel uncomfortable taking on new roles where the way they communicate is expected to more closely resemble traditional teacher norms (Liang & Matthews, 2023).

Insufficient disciplinary knowledge: A fundamental issue in the co-creation process is that some students may not possess the disciplinary knowledge needed to produce accurate and reliable content. This lack of expertise can lead to the development of educational materials that may misrepresent facts or concepts, ultimately affecting the credibility and educational value of the resources created. This issue has been widely emphasized in the computing education literature, which emphasizes the variety and complexity of misconceptions that first-year students participating in introductory courses face (Qian & Lehman, 2017). While the co-creation framework recognizes student expertise in their lived experiences without assuming they are disciplinary experts (Cook-Sather, 2014), co-creation literature also emphasizes that a lack of disciplinary knowledge requires greater scaffolding and feedback (Doyle et al., 2019). Close supervision and targeted workshops can help mitigate these concerns (Bovill et al., 2011; Lubicz-Nawrocka & Bovill, 2023). However, in large-scale educational settings with tightly packed curricula, implementing these solutions is challenging due to logistical constraints like large class sizes and limited instructional time (Bovill, 2020).

2.2 Challenges and Limitations of Previous AI Applications for Co-creation

To support the creation of learning resources, previous research used pre-transformer ML, NLP, and recommender systems. For example, collaborative filtering was employed to evaluate the quality of resources and make targeted recommendations for further review (Abdi et al., 2021; Khosravi et al., 2023). The quality of resources produced by learner-sourcing systems relied heavily on the review process. Consequently, prior studies emphasized the importance of enhancing the review process and proposed approaches to support students. This included AI assistance that highlighted specific parts of the resources to focus on when providing feedback, as well as upskilling students through short onboarding sessions and vignettes with checklists to make them more effective peer reviewers (Darvishi et al., 2022). Researchers found that students' peer review skills improved when they received both AI assistance and training. This approach was used before the widespread adoption of transformer models and was applied only during the peer review stage.

An alternative way to support students during the initial stages of creation is to provide them with automated feedback as needed. Previous research on automated feedback primarily employed rule-based algorithms and pre-transformer ML and NLP methods. These techniques were used to deliver automated feedback in the form of summative assessments based on predefined templates (Keuning et al., 2019). Pre-transformer ML and NLP methods were also used to, for instance, compare students' solutions to fully worked examples, demonstrating the correct approach and solution steps; display ML model predictions about students' expected performance; or compare students to their peers using information visualizations and learning analytics dashboards (Martinez-Maldonado et al., 2020; Di Mitri et al., 2022; Mohammadi et al., 2024). Additionally, recommendation system algorithms have been used to suggest the next learning path for students (Keuning et al., 2019; Cavalcanti et al., 2021). Despite their accuracy, these traditional AI approaches often produced uniform feedback that lacked personalization and diverse actionable steps. Moreover, they required meticulous setup and prior training on course data (Keuning et al., 2019; Cavalcanti et al., 2021). As a result, traditional AI-based automated feedback methods have been less commonly used to support students in the early stages of resource creation.

2.3 GenAI for Content Creation

GenAI presents avenues to provide the required support in the early stages of educational resource co-creation. Denny and colleagues (2023) explored whether GenAI is capable of producing educational resources and compared GenAI-generated resources with student-generated ones in the context of an introductory computer science course. They found that GenAI-generated content exhibited less variety in length and tended to mirror the provided examples. Hwang and colleagues (2024) demonstrated the feasibility of using GenAI to generate MCQs targeting varying levels of Bloom's taxonomy, while Moore and colleagues (2023) used GenAI to assess MCQs and found that traditional heuristic-based methods work better than state-of-the-art LLMs. Similarly, Jury and colleagues (2024) presented an approach using GenAI to create worked examples aiming to illustrate computer science concepts to students. Expert evaluation of GenAI-generated worked examples showed that most worked examples had clear explanations; however, they lacked depth in explanations and tended to not be optimally divided into steps (Jury et al., 2024). State-of-the-art results of GenAI related to the creation of educational resources were presented by Leiker and colleagues (2023), who used GenAI to create instructional videos with virtual tutors, and Choi and colleagues (2024), who used GenAI to create pedagogically meaningful dialogues from the pre-recorded lectures. While these studies provide nascent evidence of the feasibility and prospects of GenAI for co-creation, these works provide approaches to automate resource creation rather than aid support during the initial stages of students' co-creation.

2.4 GenAI for Feedback

Considerable effort has been dedicated to exploring the benefits of using GenAI in providing automatic feedback to students (Bernius et al., 2022; Nguyen et al., 2023; Han et al., 2023; Gombert et al., 2024; Pozdniakov et al., 2024; Hutt et al., 2024). GenAI in the form of LLMs is currently used in two main ways to provide AI feedback. The first approach, as illustrated by Bernius and colleagues (2022) and Gombert and colleagues (2024), employs LLMs to generate embeddings of student data, which are then clustered and matched with existing solutions to provide granular and customizable feedback. These works reported on students' perception of such feedback. The second approach, explored by Nguyen and colleagues (2023) and

Han and colleagues (2023), leverages the generative capabilities of LLMs to provide feedback. While the accuracy of this feedback was generally aligned with tutor-provided feedback, it sometimes contradicted the initial instruction. Despite this, the feedback was evaluated positively by both students and instructors. Lastly, recent works show avenues to use GenAI to issue feedback to support the latter stages of the co-creation loop, such as peer evaluation. Hutt and colleagues (2024) explored the avenues of using GenAI to improve peer feedback. The authors compared traditional ML methods against using LLM to evaluate peer feedback and found that although traditional methods exhibit higher accuracy, they are almost on par with LLM results (Hutt et al., 2024). Overall, the generative capabilities of LLMs show promising results, but the resulting feedback does not consistently adhere to the instructions embedded within the application.

One of the primary challenges in GenAI for tasks such as providing student feedback is the tendency of LLMs to hallucinate (Ye et al., 2023). For example, feedback generated by an LLM might include domain-specific recommendations that seem plausible and grammatically correct but are factually inaccurate. As a result, even if the feedback is specific and timely, it may lack credibility. Researchers have proposed several strategies to automatically detect hallucinations (M. Gao et al., 2024) and mitigate their risks, such as employing retrieval-augmented generation (RAG; Y. Gao et al., 2024; Li et al., 2024) or using multiple LLMs to cross-verify outputs (Gosmar & Dahl, 2025). While these approaches reduce the likelihood of hallucinations in domain-specific tasks, they come with notable limitations. For instance, RAG is not universally applicable, as it relies on a dedicated knowledge base from which relevant information must be extracted and integrated into the LLM's input. On the other hand, the method proposed by Gosmar and Dahl (2025), which involves using multiple LLMs to verify each other's outputs, shows promise but could significantly increase deployment costs, particularly when delivering AI-generated feedback to large student cohorts. However, this cost barrier may diminish over time as GenAI becomes more affordable and local LLMs improve in capability.

In order to realize the promise of AI-assisted co-creation, it is important to explore whether students take action based on AI feedback. Nascent research has indicated that the interactions between students and AI are often superficial (Shibani et al., 2024). For instance, instead of engaging with AI to improve writing coherence, students would use it to rephrase and format corrections. Singh and colleagues (2024) explored whether AI feedback could help students co-create better explanations for educational resources. They found no statistically significant differences between the quality of explanations written with and without AI support. Results of the study conducted by Darvishi and colleagues (2024) suggest that students engage with AI assistance during resource moderation, but they might depend on it rather than learn from it. The authors compared how students would engage in the moderation of resources created by fellow learners and found that students who had access to AI feedback were more productive—they identified more issues or wrote longer comments. However, once these students were asked to moderate resources without AI feedback, their moderation performance was lower than when they had AI assistance (Darvishi et al., 2024). This new evidence highlights the importance of taking into account how students interact with AI feedback and whether such feedback aids them in co-creation of educational resources.

3. Approach

3.1 The RiPPLE Platform

Deep and meaningful learning is facilitated by actively engaging with learning content, connecting and interacting with peers, and participating in activities tailored to the needs of each individual student. As workloads and class sizes increase, instructors face challenges in adopting and consistently implementing these strategies. In response, we created RiPPLE (Khosravi et al., 2019), based on the belief that every student can create impactful waves in the ocean of knowledge. Instead of being passive recipients of content, RiPPLE empowers co-creation at scale (Dollinger et al., 2024) by enabling students to share their wisdom and contribute to knowledge creation, helping foster active, social, and personalized learning experiences.

Figure 1 provides an overview of the three main learning activities in RiPPLE: create, evaluate, and practise. Initially, students create educational resources. These are then evaluated through a peer-evaluation process where students assess the effectiveness of the content. Effective resources are approved and added to a repository accessible to all students in the course, while ineffective resources are sent back to the author for re-submission. As students practise with approved resources, RiPPLE's AI algorithm models each student's knowledge level on course topics. It then uses this information within its adaptive engine to recommend personalized learning resources based on individual learning needs. Additionally, RiPPLE employs analytics and AI spot-checking algorithms to provide academic oversight of the student content creation process, minimizing the workload for instructors. This inspection model sends weekly insight notifications, highlighting resources that would benefit most from expert judgment. The platform has been widely used, with over 60,000 students using RiPPLE in more than 250 course offerings across various disciplines. Currently, RiPPLE allows authoring educational resources of various types, such as MCQs with one or multiple answers, long answers, worked examples, short notes, research reports, and explanations. This study focuses on aiding students during the first step of the co-creation process depicted in Figure 1.

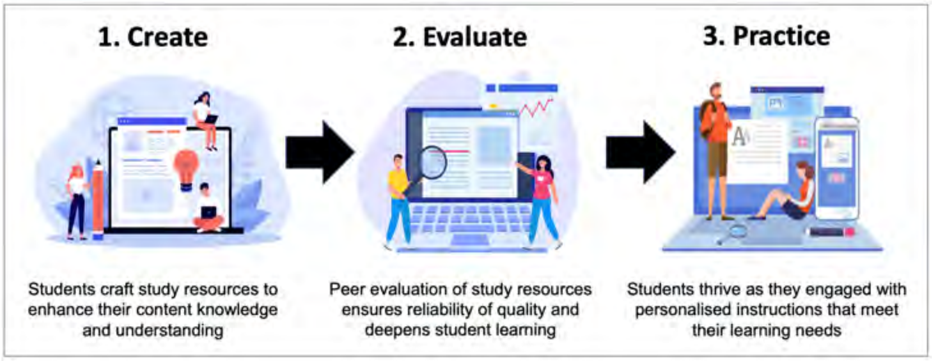

Figure 1. Basic overview of the three main interconnected learning activities in RiPPLE.

3.2 Addition of AI Assistance to RiPPLE

To help students create educational resources, we decided to implement AI assistance. The AI assistance was available to students during the creation phase. Once a draft of the educational resource was completed, students clicked a button to receive AI feedback. The AI assistance is implemented as on-demand feedback available in a sidebar, as shown in Figure 2. Typical AI feedback included three sections: a summary of the aims of the current educational resource the student created; *positive appraisals*, emphasizing the strengths of the resource; and *suggestions for improvement*, advising students on how to improve the resource's quality (see Figure 2, for example). When students asked for the AI feedback, the current state of the resource they worked on was used to fill in prompts and prepare a query for GenAI.

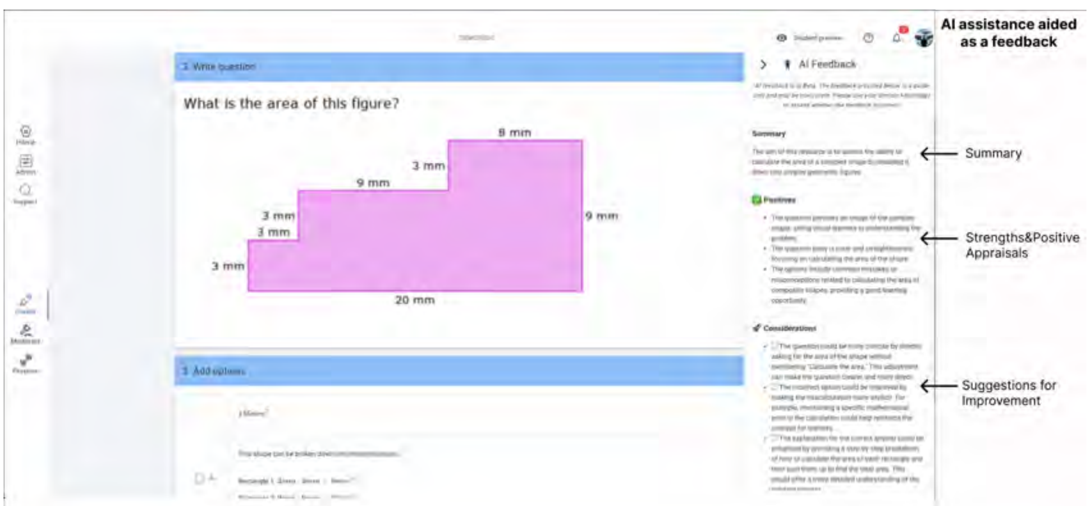

Figure 2. The AI assistance in the form of instant feedback implemented during the *creation* phase in RiPPLE.

To implement the AI assistance, we followed steps 1 to 4 from the framework for pedagogical incorporation of GenAI for educational purposes, as suggested by Pozdniakov and colleagues (2024). Our approach was as follows. We began by defining the aim of the GenAI application: to support students in the initial stages of co-creation. Specifically, our goal was to help students overcome knowledge and skills gaps identified in the background, such as a lack of higher-order cognitive skills, insufficient pedagogical expertise and experience, and limited disciplinary knowledge.

Next, we consulted the feedback design framework, which emphasizes the importance of including two critical aspects in feedback: identifying the current state ("What did I do well?") and recommending actionable next steps ("Where to go from here"; Hattie & Timperley, 2007). We then defined the desired tone for the response, using instructions such as "You are an expert exam question writer and tutor," and provided examples illustrating the expected tone, such as "Your feedback must be kind, constructive, specific, and very actionable," following recommendations from Mollick and Mollick (2023). We did not systematically verify whether these recommendations were adhered to in the final feedback prior to deployment, which is why RQ1 focuses on evaluating the scope, depth, and tone of the generated feedback.

The final step outlined by Pozdniakov and colleagues (2024) involves selecting explicit criteria to guide LLMs in generating output. Drawing on Collins (2006), we identified three essential criteria for effective MCQ design: *quality of the question's stem*, *quality of options*, and *quality of explanation*. Since we used the LLM in the deployed version of the tool, we included instructions to ensure that the prompt would generate feedback formatted within HTML tags. This was necessary to structure the feedback into three distinct sections: summary, strengths and positives, and suggestions for improvement.

We conducted ad hoc experimentation to determine the optimal prompt that consistently met both feedback quality and formatting requirements. To ensure that the feedback was well structured and aligned with our criteria, we tested zero-shot and few-shot prompting techniques (1 and 2 shots), ultimately choosing a 2-shot example prompt without reasoning (R. Wang et al., 2023), as detailed in Appendix C, Listing 2. Before finalizing the k-shot prompt, researchers manually evaluated two generated feedback examples (later included in the prompt shown in Appendix C, Listing 2) to confirm that they adhered to the three feedback quality criteria and were correctly formatted using HTML tags. While we made slight adjustments to the prompt during data collection, all major elements remained intact. In Appendix C, Listings 1 and 2 provide the final prompt used with OpenAI's GPT-3.5-turbo model. We selected a temperature value near zero to minimize inconsistency, balancing cost and quality for scalable feedback generation.

4. Evaluation Approach

In this section, we outline the methodologies and analytical techniques employed to assess the AI feedback provided to students. This includes a detailed description of participants and data collection, the characteristics and depth of the AI feedback, and the impact of the AI feedback on student authoring activities, as well as an analysis of students' perceptions of the AI feedback. The overall analysis procedure is shown in Figure 3.

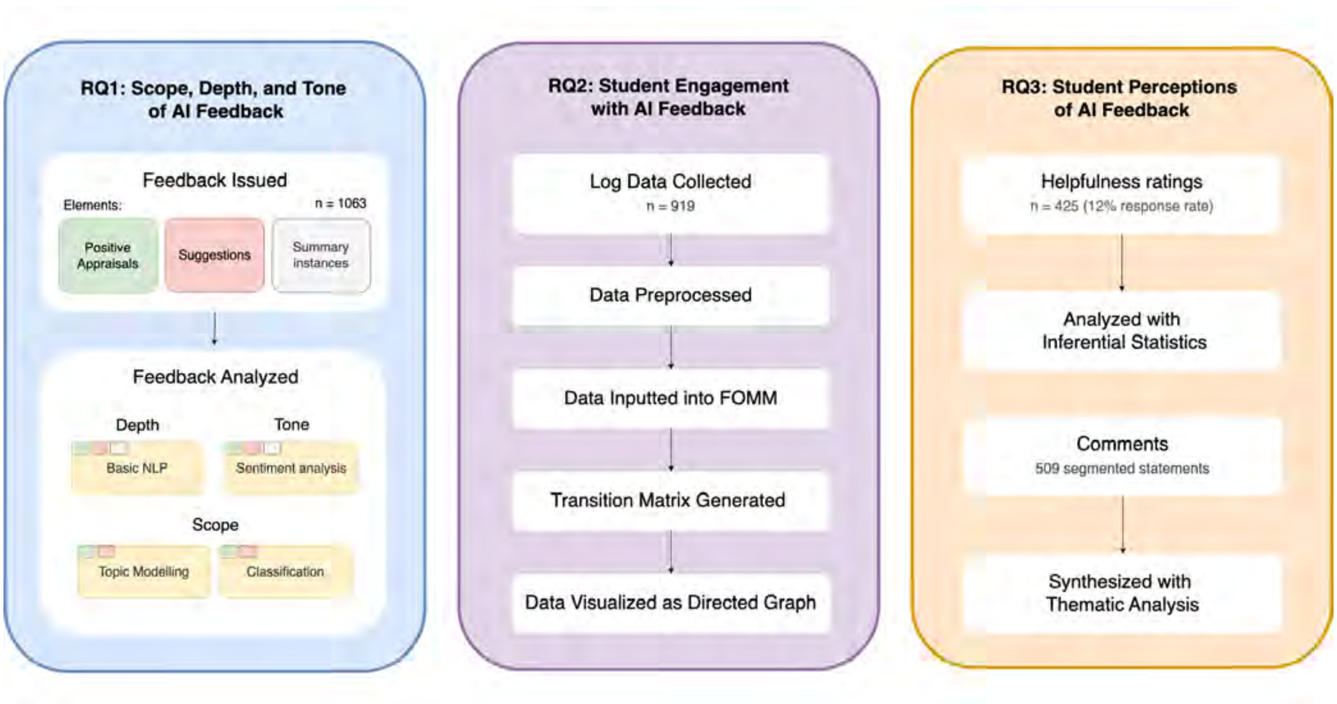

Figure 3. Analysis employed in the current work (FOMM = first-order Markov model).

4.1 Participants and Data Collection

A total of 3324 students from an Australian university who used RiPPLE in their course consented to participate in this study from July 2023 to June 2024, after approval by an institutional ethical review committee. The data used in the current study includes students' logs and the AI feedback captured when students worked on MCQ resources. The reason for this is that this type of question has an equal number of elements; hence, regardless of disciplinary differences, the log-capturing software and the AI feedback would target the same elements. The data collected was from multiple introductory courses, including information systems, economics, and health sciences.

4.2 Learning Context

Consented students used RiPPLE to co-create MCQ resources during multiple assessment, where their work contributed 10% to their final grade. The MCQs consisted of several parts: a title; question details such as its topic and difficulty; the question's main body (stem), which introduces the question and its context; correct and incorrect options (distractors), among which students were required to choose only one correct answer; and explanations for both correct and incorrect options to facilitate learning.

The AI assistance was available to students on demand. A typical AI-assisted resource-editing session looked as follows. Students started their work on the resource, and once students filled in all the MCQ sections, they asked for AI assistance. After that, students were encouraged to rate the helpfulness of the AI feedback and provide short comments in the sidebar. Next, students could either consider the AI feedback received and improve the quality of the resource or proceed to the self-assessment, where they were asked to rate the quality of their work and the confidence in their judgment. Importantly, after finishing the self-assessment, students could still make changes to the resource. Finally, students proceeded to the resource submission.

4.3 Data Analysis and Measures

4.3.1 RQ1. Analysis of AI Feedback

The data and analysis for this RQ are as follows. In total, 1063 AI feedback instances were analyzed. This feedback was stored as text consisting of three sections: resource content summary, positive appraisals, and suggestions for improvement. Each of these sections typically consisted of multiple sentences, typically represented in text using a bullet-pointed list, each providing a conclusive appraisal or suggestion. We used each individual feedback element presented as a bullet point as a unit of analysis. In total, we had 776 instances of the summary category, 3054 positive appraisals, and 3317 suggestions for improvement. A typical AI feedback instance included three positive appraisals and three suggestions for improvement. The AI feedback issued to students was then examined through the lenses of the three aspects described below.

Scope of the AI feedback. To analyze the scope, we conducted a thematic analysis of a range of strengths and suggestions included in the feedback in two steps. First, we used topic modelling to extract thematic categories from the feedback. Second, we used LLMs to classify the feedback given the thematic categories informed by the topic modelling. We included only positive appraisals and suggestions for improvement, as the summary section of the feedback was not designed to provide meaningful feedback on the resource.

We used topic modelling to overview the scope of the generated feedback. Topic modelling identifies semantically similar themes in text data, useful when manual inspection is impractical (Y. Chen et al., 2023). We used the BERTopic package (Grootendorst, 2022) with the “fasttext” model for short texts. To do this, we extracted embeddings from data, transforming input into high-dimensional vectors; applied UMAP reduction to preserve structure while reducing dimensionality; and then used HDBSCAN to cluster reduced embeddings into topics (Grootendorst, 2022). Following that, we computed class-based word frequency (cTF-IDF) for human-interpretable topic representation. We used uni-grams and bi-grams with a minimum frequency of four to represent each topic. We applied topic modelling separately for positive appraisals and improvement suggestions, resulting in two models. We experimented with 20, 15, and 10 topics, choosing 10 for both models to keep them content-agnostic, i.e., not focusing explicitly on the course subject, yet including prominent feedback characteristics. Distribution-based outlier reduction attributed all but five appraisals to one of the 10 topics. We manually grouped semantically similar topics into three categories: “1-Clarity&Writing,” “2-MCQ Design,” and “3-Disciplinary Knowledge.” The resulting topics can be found in Appendix A, Tables 4 and 5.

To classify each feedback item within the three defined categories, allowing for simultaneous membership in multiple categories, we decided to use LLMs. This decision was based on the limitations of topic modelling, which, while effective for exploratory analysis, is less reliable for classification tasks involving overlapping categories (Y. Chen et al., 2023). To obtain more precise estimates of the scope of AI feedback, we turned to LLMs. However, emerging research highlights that LLMs often struggle with consistent performance in text classification, particularly when dealing with complex topics (R. Wang et al., 2023; Zambrano et al., 2023; Misiejuk et al., 2024). For instance, R. Wang and colleagues (2023) demonstrated that using 3-shot examples with reasoning prompts achieves greater alignment with human annotations when classifying comments from students watching lecture recordings. Another study examining the alignment of GPT models with human annotations for student-generated content in online discussions found that premier models like GPT-4, when used with 1-shot examples, generally performed well but sometimes failed to detect false negatives across several codes (Misiejuk et al., 2024).

These studies underscore the difficulty of ensuring high reliability with LLMs, particularly in complex and context-dependent coding tasks where subtle nuances can be overlooked. Furthermore, the opaque nature of LLM decision-making processes complicates efforts to understand why a specific classification or output was generated. This stands in contrast to traditional NLP methods, which can be paired with explainability techniques like local interpretable model-agnostic explanations (LIMEs) to provide clearer insights (Hutt et al., 2024; Huang et al., 2025). Recent findings by Gosmar and Dahl (2025) indicate that using multiple LLMs in a sequential pipeline—where each LLM can access the results of previous inferences—reduces the rate of speculative or inaccurate outputs and enhances the accuracy of reasoning. Our approach aligns conceptually with that of Gosmar and Dahl (2025) but with two key distinctions. First, in our framework the initial two LLMs do not have access to each other’s outputs. Second, while Gosmar and Dahl (2025) used LLMs from the same model family (e.g., ChatGPT), research suggests that LLMs may exhibit bias toward their own outputs or those from the same model family (Panickssery et al., 2024), potentially compromising reliability. To address this, we opted to use models from different architectural families, thereby reducing the risk of such biases.

In light of this evidence on effective prompting techniques for improved classification accuracy, we adopted the following approach. We incorporated 3-shot examples with reasoning into our prompts, as recommended by R. Wang and colleagues (2023), to enhance classification quality. We used two models—the open-sourced "Llama3.1-70b," deployed locally on university clusters, and "GPT-4o-mini" (November 2024)—to classify the data. We then calculated weighted Cohen's k to measure the agreement between the two models for each category. This analysis showed moderate agreement for the category "1-Clarity&Writing" ($k = 0.55$), high agreement for "2-MCQ Design" ($k = 0.81$), and substantial agreement for "3-Disciplinary Knowledge" ($k = 0.72$). To address instances of disagreement between the models, we employed a third model, "Claude 3.5 sonnet version 2" (November 2024), to make final classification decisions and reconcile discrepancies. Following that, we visualized the distributions of thematic categories and plotted them using a Venn diagram. Due to parsing issues of the outputs provided by LLM, we ended up with 2935 positive appraisals and 3233 suggestions for improvement classified, which explains the differences between the total number reported earlier in the section.

Although this multi-agent workflow is designed to enhance robustness, it is essential to recognize that relying exclusively on LLMs for final classification decisions can still pose challenges. This is due to their potential biases, limited contextual understanding, and inherent reasoning constraints. For example, findings from Xu and colleagues (2024) revealed that while the agreement between human coders and LLMs was moderate at the level of code categories, agreement at the individual code level was significantly lower. This highlights the ongoing need for human oversight, particularly during the evaluation phase (Xu et al., 2024).

Depth of the AI feedback. We focused on the complexity of the generated feedback across three structural levels: word, sentence, and overall text. At the word level, we used word count (WC) and complex word count (CWC) metrics to evaluate lexical complexity. WC measures feedback length, revealing content volume. CWC assesses vocabulary sophistication, tagging words with more than two syllables as complex. A higher CWC indicates advanced vocabulary, impacting readability and comprehensibility. Combining WC and CWC provides insights into language difficulty. At the sentence level, we used average sentence length (ASL) and proportion of advanced sentences (POAS) to highlight syntactic complexity. ASL is the average number of words per sentence, with longer sentences indicating more complex structures. POAS, using the BERT model fine-tuned on the OneStopEnglish (OSE) dataset (Liu & Lee, 2023), was used to classify sentences into beginner, intermediate, and advanced levels. We calculated the proportion of advanced sentences in each feedback instance, with a higher POAS indicating greater difficulty for students (Martinc et al., 2021). At the Overall Text level, we evaluated holistic complexity using Flesch Reading Ease (FRE; Flesch, 1948) and the Crowdsourced Algorithm of Reading Comprehension (CAREC; Crossley et al., 2019). FRE combines ASL and average syllables per word to assess readability, with scores ranging from 0 (unreadable) to 100 (very easy). CAREC uses advanced NLP tools and crowd-sourced judgments to evaluate text comprehensibility, encompassing 13 linguistic features related to lexical sophistication, syntactic complexity, and text cohesion. A higher CAREC score indicates more difficult text. We used NLP tools from the Georgia State University website for evaluation[1].

Tone of the AI feedback. Finally, we evaluated the tone of the generated feedback using sentiment analysis. Specifically, we employed VADER (Hutto & Gilbert, 2014), a fast rule-based model, which uses a predefined dictionary of sentiment-related words and rules to assess the emotional quality and intensity of the feedback, similar to the process of Misiejuk and colleagues (2021). VADER categorizes feedback into positive (scored by 1), negative (scored by -1), and neutral (scored by 0) tones and provides sentiment scores to reflect the strength of these emotions. Positive feedback, indicated by higher positive sentiment scores, is presumed to foster greater motivation and engagement than negative or neutral tones.

4.3.2 RQ2. Analysis of Student Engagement with the AI Feedback

To analyze student engagement with the AI Feedback, we examined log data from 919 students co-creating with AI who consented to participate in the study. The log data included students' resource editing activity recorded only after AI feedback was issued. The data processing included filtering all the observations prior to receiving feedback and sorting the student resource editing logs with respect to the time they were performed. Students performed a total of 2270 resource editing activities, where a typical student would perform a single editing activity after receiving the AI assistance ($Mdn = 1$). These resource activities were editing question details, question bodies, or question options. The resulting log of a sequence of resource editing activities was used as input for the first-order Markov model (FOMM), which was done using the `pMineR` package (Gatta et al., 2017). FOMM outputs a transition matrix indicating the probabilities of switching from one editing activity to another. We then used this transition matrix to visualize FOMM as a directed graph. In the graph, each node corresponds to a learning editing activity students did, while the edge indicates the transition from one type of editing activity

[1]Access Date: 20 October 2024, https://nlp.gsu.edu/home

to another. The graph is directed and each edge has a probability, indicating the likelihood of doing one editing activity or another. This allowed us to analyze how students acted based on the AI feedback. The preprocessing and analysis for this RQ were done similarly to how it is reported in Lahza and colleagues (2022) and Fan and colleagues (2022).

4.3.3 RQ3. Analysis of Students' Perception of the AI Feedback

To analyze students' perception of the AI feedback, we collected evaluations from 3324 students who received the AI feedback while creating MCQ resources. After students received the AI feedback, they were asked to rate how helpful it was on a 5-star scale. Students were also asked to provide comments on the feedback's helpfulness. We conducted a chi-square goodness-of-fit test to assess whether the distribution of helpfulness ratings—negative (1–2), neutral (3), and positive (4–5)—was uniform. All assumptions for the chi-square test were met. The analysis of students' comments began with data cleaning, excluding inapplicable or ambiguous responses, such as "Yes." Next, the second author used an inductive approach in NVivo to code the remaining data, allowing themes to emerge from the data (Saldaña, 2015). The codes and themes were then reviewed and refined through discussions with the research team. Although attempts were made to further subcode the theme of general valuable aspects, limited consensus among the research team prevented additional categorization. Each unit of analysis consisted of a complete or partial comment (e.g., "It provided great guidance for improving the question..."). Out of all participating students, 13% ($N = 425$) submitted comments, yielding a total of 509 coded statements. The coding scheme is available in Appendix B, Table 6.

5. Results

5.1 RQ1: The AI Feedback

Scope of the AI feedback. The thematic category "1-Clarity&Writing" included elements of AI feedback, which focused on the aspects of students' phrasing and writing all parts of the resources. A characteristic of this theme is that both positive appraisals and suggestions present in the feedback primarily focused on the presentation of how resources are written (e.g., whether writing in options, explanations, or the main question body reads well and is grammatically correct). An example of a positive suggestion from this thematic category is "The question body is clear and grammatically correct." A typical suggestion for improvement is "Perhaps consider breaking down the definition into smaller parts for better clarity and easier absorption of information." In the case of positive appraisals, this category might assess whether parts of resources are concise (see Table 3, "Positive appraisals examples" column in Appendix A). In case of suggestions for improvement (Table 3, "Suggestions examples" column), such feedback elements might include comments regarding how to improve the written presentation of the resource description, suggesting, for instance, "[breaking the explanations] into smaller parts for better clarity and easier absorption of information."

The thematic category "2-MCQ Design" consisted of feedback on the effectiveness of the design of the MCQ for learning relating to either the whole resource or its parts, such as the option set. This thematic category typically involved assessments of the logical plausibility of incorrect options, evaluating how straightforward it was to pick up incorrect options, and whether the phrasing of the explanation is comprehensible. For instance, a typical positive appraisal in this category might depict whether an incorrect option is indeed factually incorrect, for example, "The options are plausible and factually correct." Suggestions for improvement typically follow this pattern, with the only difference being that they might provide content-specific suggestions on how to improve the MCQ item. For instance, the following suggestion is representative of this thematic category: "Explanation 3 should clearly explain why the option is incorrect. For example: Incorrect. The option includes the attribute 'orderNum' as underlined, but the SQL statement does not mention underlining any attribute."

The thematic category "3-Disciplinary Content" included feedback elements primarily focusing on the recommendations about course-specific knowledge that a resource targets. The feedback elements from this category either praise or suggest whether explanations of the options contain enough context to understand the subject matter and, if not, suggest improving the explanations; they also evaluate whether the most common subject-specific knowledge is included in the resource. A typical positive appraisal belonging to this category looked like this: "The suggestion for creating a structured framework to avoid entrepreneurial pitfalls is a positive contribution." A typical suggestion for improvement would look similar but include advice for improvement: "It may be helpful to delve deeper into the competitive landscape and how Polarity plans to differentiate itself from potential competitors." Full results of the topic modelling are presented in Appendix A, Tables 3, 4, and 5.

We also found that feedback elements that belong to a single thematic category show slight differences between positive appraisals and suggestions for improvement, as shown in Figure 4. Most single-category feedback focuses on assessing "3-Disciplinary Content" (Figure 4, 28% for positive appraisals and 23% for suggestions for improvement), followed by a much smaller emphasis on "2-MCQ Design Principles" (8% and 4%) and "1-Clarity&writing" (8% and 2%). Importantly, there are notable similarities and differences in how positive appraisals and suggestions for improvement address feedback elements spanning multiple categories (see Table 1). Both types of feedback frequently highlight strengths in adhering to best practices for MCQ design and disciplinary content, with similar proportions (34% and 33%, respectively). However, a key

Table 1. Feedback elements belonging to more than one thematic category. The first column indicates the intersection of thematic categories; the second and third columns present examples of such intersections for appraisals and suggestions for improvement.

Thematic Category	Positive Appraisal—Representative Feedback Element	Suggestions for Improvement—Representative Feedback Element
MCQ design ∩ Disciplinary content ∩ Clarity & Writing	The correct answer is logically sound, explaining the concepts of absolute and comparative advantage accurately.	The options should avoid clues that give away the correct answer. For instance, option 4, All of the above, could be misleading as it includes all the previous options, potentially giving away the answer. This could be revised to make it less obvious. The explanation for option 2 may need to be clarified. While interactive use is an advantage, stating that it is "very interactive" might be too subjective and vague.
Disiplinary content ∩ MCQ design	The incorrect options are factually inaccurate and help us to distinguish between the correct and incorrect applications of the Gini coefficient.	The question could be more meaningful without having to read all the options first, for example, rephrasing the question to focus on a specific aspect of SQL's advantage, such as Which of the following statements best describes an advantage of SQL?
MCQ design ∩ Clarity & Writing	The explanation provided for the correct option is concise and accurate, reinforcing why the chosen format is the right answer. It helps clarify the reasoning behind the correct choice effectively.	Consider rephrasing the question stem to make it more meaningful without having to read all the options first. This can enhance the clarity and efficiency of the question.
Disciplinary content ∩ Clarity & Writing	The explanations are accurate and provide clarity on why each option is considered an advantage of SQL.	The question body could be more concise by removing the unnecessary repetition of the GDP and trade volume values in each option. For example, instead of restating the values in each option, the question could simply ask, What is the trade openness of a country with a trade volume of 200 million and a GDP of 100 million? This revision would make the question clearer and more focused.

Figure 4. The Venn diagrams illustrate the distribution of feedback categories for positive appraisals (left) and suggestions for improvement (right) across three dimensions: Clarity & Writing, MCQ Design, and Disciplinary Content. Only a small portion of positive appraisals encompassed all three thematic categories, while a more substantial share of suggestions for improvement addressed multiple aspects of resource quality, showing greater versatility.

distinction lies in comprehensiveness: only a small fraction of positive appraisals incorporated all three thematic categories, while a significant portion of suggestions for improvement addressed various aspects of resource quality more comprehensively. The key difference is that only 4% of positive appraisals incorporated all three thematic categories, while 20% of suggestions for improvement appeared to be versatile when making suggestions on different aspects of the resource quality. This suggests that the AI feedback tends to highlight student strengths by emphasizing a combination of two thematic categories—such as clarity and linguistic consistency with disciplinary content or MCQ design with disciplinary content. In contrast, suggestions for improvement more often identified and addressed shortcomings across all three thematic areas included in the resources.

Depth of the AI feedback. In contrast to the scope of feedback described above, the analysis of the depth of the AI feedback included reporting descriptive statistics for summaries. A typical AI feedback instance consisted of 180 words, whereas each sentence contains around 16 words, with a median readability score of .34, indicating that they should be easy to understand.

Table 2 and Figure 5 present the aggregated word, sentence, and feedback text-level metrics. First, it can be seen that summaries are the least verbose feedback category, in terms of both total and complex words included. It is followed by positive appraisals, which have a median word count of 4, while the suggestions have more than twice as many words as positive appraisals do (see Table 2, columns WC, CWC). At the same time, a typical summary included larger and more complex sentences, while positive appraisals and suggestions for improvement were shorter and less complex (see Table 2, columns ASL and POAS, $Mdn_{\text{ASL\_summary}} = 20$, $Mdn_{\text{ASL\_positives}} = 15.3$, $Mdn_{\text{ASL\_suggestions}} = 16.4$). When it comes to the overall text metrics, FRE scores indicate that the summary was the easiest feedback category to read ($Mdn_{\text{FRE\_summary}} = 52$), while positive appraisals were the hardest, with suggestions for improvement in between ($Mdn_{\text{FRE\_positives}} = 38$ and $Mdn_{\text{FRE\_suggestions}} = 45$). However, CAREC shows that summary and positive appraisals have almost the same level of readability ($Mdn = .21$), while suggestions for improvement might have been harder to read ($Mdn = .21$).

Table 2. Summary statistics for feedback categories (Median/SD).

Categories	Word-Level Metrics		Sentence-Level Metrics		Overall Text-Level Metrics	
	WC	CWC	ASL	POAS	FRE	CAREC
Summary	21/6	3/1.7	20/4.5	1/0.45	52/13.6	0.21/0.11
Positives	43/15.2	9/4.6	15.3/4.7	0.6/0.39	38/13.3	0.21/ 0.07
Suggestions	117/36	23/7.9	16.4/2.5	0.6/0.22	45/9	0.3/ 0.07
Overall	180/43.2	36/9.95	16.3/2.4	0.6/0.23	44.4/9.4	0.34/0.06

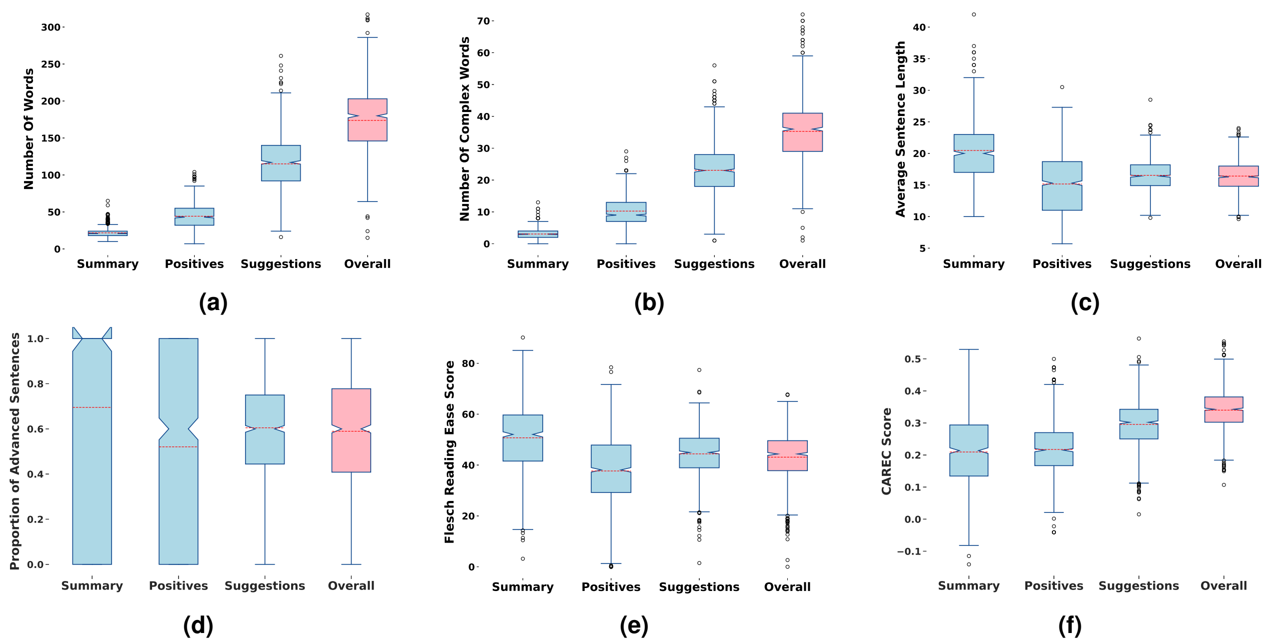

Figure 5. Distribution of the text complexity scores, including (a) NW, (b) NCW (words with more than two syllables), (c) ASL, (d) PAS, (e) FRE, and (f) CAREC, for GenAI feedback across *summary*, *positives*, *suggestions*, and *overall* text categories.

Tone of the AI feedback. Figure 6 presents the distributions of VADER scores resulting from sentiment analysis. Overall, the median VADER score for the AI feedback equals 0.94, which indicates that the AI feedback had a positive sentiment. The median VADER score for summaries equals 0, which indicates a neutral sentiment. Positive appraisals had a median score of 0.68, while for suggestions for improvement, the score is approximately 0.84. As such, both of these feedback elements have positive sentiments. Importantly, VADER scores are uniformly distributed when it comes to summaries and suggestions for improvement, while the distribution for positive appraisals is skewed more toward upper limits.

5.2 RQ2: Student Engagement with the AI Feedback

In total, students asked for the AI assistance 919 times. After the feedback was issued, 336 students made corrections to the MCQ options ($N = 270$); question body ($N = 217$); and question details such as difficulty and topic ($N = 49$) and completed self-assessment ($N = 867$), which preceded the resource submission. Figure 7 represents how students proceeded with resource creation once they received the AI feedback. It indicates that after receiving the feedback, in 65% of resource creation sessions,

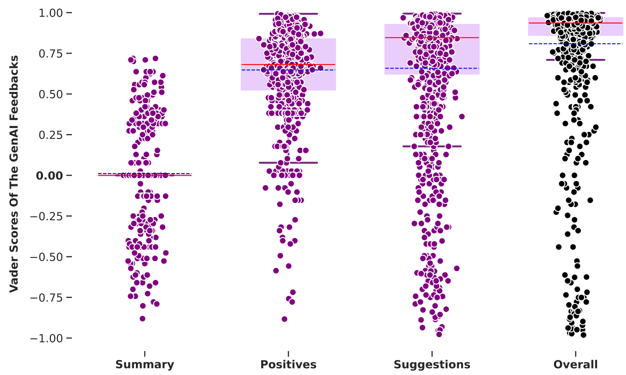

Figure 6. Distribution of the Vader sentiment scores for the AI feedback across *summary*, *positives*, *suggestions*, and *overall* text categories. The sentiment scores range from -1 (most negative) to 1 (most positive), with a score of 0 indicating neutral sentiment. Each dot represents an individual feedback score. The red lines represent the median sentiment scores, while the blue dashed lines represent the mean sentiment scores. The shaded areas indicate the interquartile range for each category.

students were satisfied with the quality of their resource and proceeded to self-assessment (AI $\rightarrow$ SE)[2]. Following this, in 88% of creation sessions, students proceeded to the submission (SE $\rightarrow$ Sub). Notably, if students decided to do additional work on the resource after they received the feedback, the sequence of their editing activities would look as follows. If students decided to proceed to work on the question body (AI $\rightarrow$ Q: 16%), they would typically end up with some follow-up edits, with the probability of transition to self-assessment or submission being lower compared to the rest of the resource elements edits (Q $\rightarrow$ SE: 42%, while Opt $\rightarrow$ SE and QD $\rightarrow$ SE at 51% were higher). These edits would almost certainly be refining the options (Q $\rightarrow$ O: 42%). Alternatively, if students decided to modify options (AI $\rightarrow$ O: 12%) after receiving the AI assistance, they would likely proceed to self-assessment after that (O $\rightarrow$ SE: 51%). Students rarely decided to modify the question details based on the AI feedback (AI $\rightarrow$ QD: 3%). The most prominent follow-up activity after making AI-informed modifications to the resource is self-assessment, which happened in nearly 50% of the resource editing situations (Q $\rightarrow$ SE: 42%, Q $\rightarrow$ SE: 51%, and O $\rightarrow$ SE: 56%). Once self-assessment was initiated, students could make additional modifications to the resource, which was represented by directed connections from SE to the resource elements and direct connection to the submission. However, these events were unlikely to happen in less than 7% of resource editing sessions.

Overall, these results show that in the majority of situations students submitted their work once they had received the feedback without any edits. However, if students decided to improve the resource after receiving AI feedback, they tended to either work on the question body, which would necessitate a chain of further changes, or make minor changes to the question options or details, e.g., specify the topic or question difficulty, and proceed to the self-assessment and submission.

5.3 RQ3: Student Perception of the AI Feedback

Results indicated that 79% of students rated the helpfulness of the AI feedback as positive, 12% as neutral, and 9% as negative (see Figure 8). This suggests a strong preference among students for rating the feedback as helpful, $\chi^2(2, N = 3324) = 3149$, $p < .001$. To explore the reasoning behind students' ratings, we thematically analyzed their comments about the feedback. We found that of the 509 coded data units from the comments $N = 386$ stated that the AI feedback was "valuable" in a nonspecific or uncharacteristic way. For instance, students described the feedback as "helpful" and "good," with one student stating: "Wow this is amazing and super helpful, I'm literally blown away." Students often referred to the AI assistance as being "clear," "comprehensive," "kind," or "reasonable." Students also mentioned that the AI assistance helped them to improve the disciplinary content of their resource, for example, "It also allowed me to see that my explanation should explain why it is the best answer and how to relate it back to the concept further." Numerous comments also mentioned that the AI-provided suggestions were valuable ($N = 32$; e.g., "The suggestions are helpful"), with an additional 49 indicating that the feedback helped improve students' answers or distractors ($N = 8$); explanations ($N = 8$); question body ($N = 13$); or other aspects, like the resource's title or difficulty ($N = 4$). A few coded statements referred to the AI assistance helping improve clarity and writing ($N = 16$; see Appendix B, Table 6, for the full coding scheme).

When students commented on issues and challenges of the AI feedback ($N = 123$), they commonly expressed that the feedback was inaccurate ($N = 31$) or provided redundant information ($N = 26$), such as suggestions that were already implemented. Sixteen times, students expressed disagreement with the feedback, for example: "I don't know if all of the 7

[2] The arrow symbol means a transition from activity A to activity B.

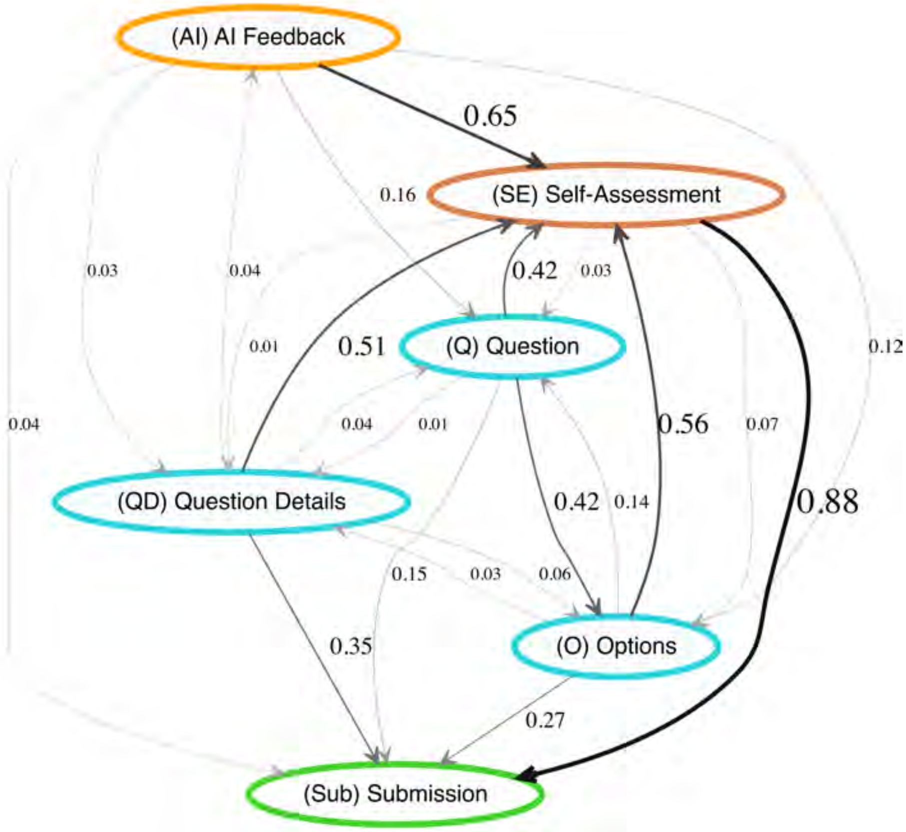

Figure 7. Overview of the resource editing activity after students asked for the AI assistance. The nodes represent the resource editing activity students performed, while the directed edges represent the editing activity sequence students took. The thickness of the edge denotes the probability, also given as a number next to the edge, of taking on the next editing activity, such as working on question phrasing, improving MCQ options, or refining question details. Typically, students made progress on the resource before asking for the AI assistance. After that, students would work on the main MCQ body ("question"), rethink resource difficulty and expected topic ("question details"), or revisit phrasing for options and their explanations ("options"). "Self-Assessment" depicts students evaluating the quality of the resulting resource and their confidence in it and would typically precede the resource submission. After submission, the resource was sent for peer review.

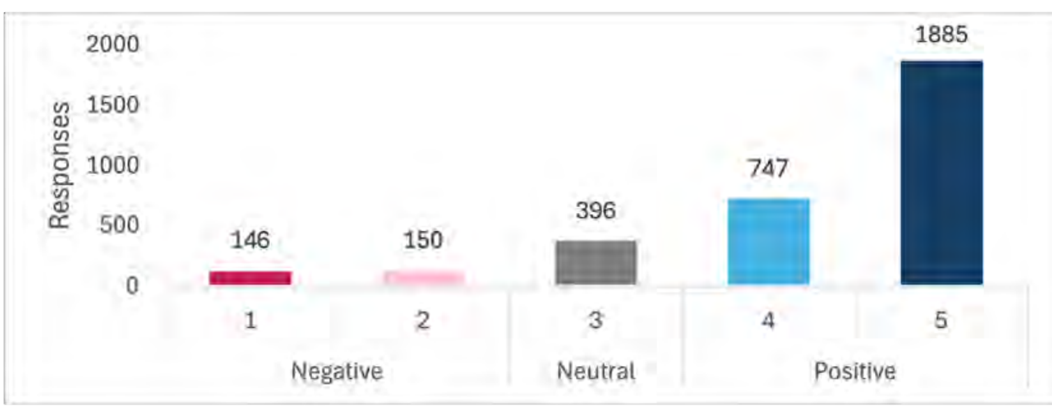

Figure 8. Distribution of student ratings for the helpfulness of the AI assistance on a scale from 1 (not helpful) to 5 (very helpful).

would actually make it better." Students also referred to the AI assistance as being "unclear" ($N = 11$) and misunderstanding the author's aims or the task requirements ($N = 13$). The rest of the coded data included students' perceptions of the AI assistance as having various uncategorizable issues, such as challenges with feedback representation, repetitiveness, and practicality ($N = 11$). Lastly, some comments suggested that the AI feedback was not helpful ($N = 15$), without providing a reason; for example, "It was just fine"; see Table 6 in Appendix B.

6. Discussion and Lessons Learned

This work has made the following contributions. Our main contribution is an approach to support students' co-creation of educational resources, overcoming a persistent concern that students' delivery of high-quality educational resources could be impeded by a lack of higher-order cognitive skills, pedagogical experience, and disciplinary knowledge (Doyle et al., 2019; Darvishi et al., 2021; Moore et al., 2021; Khosravi et al., 2023). First, aligned with the results of previous studies (Dai et al., 2023; Denny et al., 2023; Kazemitabaar et al., 2024), our results suggest that GenAI could be instrumental in producing concise, structured, and specific feedback for students, which students tend to perceive as helpful. Second, our findings indicate a relatively high actionability rate. Based on our results, we suggest differentiating between AI assistance during the ideation and refinement stages of resource editing, thus contributing by illustrating two strategies for how AI assistance could be implemented to support students' actionability based on AI feedback during the resource editing. During the *ideation stage*, where students think about the ideas for the resource, AI can assist by providing inspiration, suggesting ideas, or providing initial drafts or prototypes. This can help aid students' creativity. During the *refinement stage*, where students' ideas need to be polished and developed, AI can assist by identifying areas for improvement, suggesting enhancements, or providing assessments of the work. Lastly, the results of our study contribute to the existing literature on feedback perception by confirming that students tend to find our implementation of AI feedback helpful when creating MCQs; however, perceived helpfulness could deteriorate when feedback is seen as inaccurate or inconsistent. We contribute to the discussion on how to resolve these issues.

6.1 Findings from the Research Questions

RQ1: What specific feedback did the AI assistance provide to students? Our results suggest that the AI feedback issued to students could incorporate discipline-related knowledge and MCQ-specific suggestions while changing the vocabulary complexity where appropriate and maintaining a positive sentiment. This means that such AI feedback includes recommendations regarding course-specific knowledge gaps students might expose when they engage in resource authoring, but such AI feedback also provides detailed instructions about which parts of MCQ resources should be improved. However, our results also emphasize the difficulty of designing versatile AI assistance that would consistently combine all the required feedback categories and follow the instructions flawlessly. Our results are aligned with nascent studies showing that GenAI could be applied to generate lexically varied and logically sound feedback based on students' inputs (Dai et al., 2023). Importantly, our results demonstrate that the AI feedback is concise and typically consists of 180 words, i.e., no longer than a typical X post. These results should be seen in light of previous research emphasizing that the feedback presentation is essential for it to be read and acted upon by students (Panadero & Lipnevich, 2022).

RQ2: How did students engage with and use the feedback provided by the AI assistance? Our findings indicate that students typically would attempt to act based on the AI feedback in less than half (35%) of the resource editing sessions. This result might indicate that students tended to ask for AI assistance once they had already done substantial work, which could be ready for submission. Previous studies indicated that only a minority of students directly acted on the suggestions provided in the feedback messages as captured by "calls to action" (Iraj et al., 2021); similarly, only a minority of feedback issued via Learning Dashboards is directly acted upon (Matcha et al., 2020). Importantly, while previous research found a positive relationship between engagement with feedback and increased success rate in the course (Pardo et al., 2019; Iraj et al., 2021), these studies neither log nor report how students act based on the feedback, which makes it difficult to compare our results against previous research. As such, our results might indicate that students demonstrated high feedback actionability. Importantly, our results should be seen in light of the results reported in a randomized controlled field experiment conducted on the same platform (Darvishi et al., 2024), which showed that in a condition with AI assistance, students made more changes compared with the condition without AI assistance.

Second, the AI feedback used in this study was delivered to students in a separate sidebar window, as opposed to, for instance, writing tools (Knight et al., 2020) or integrated environments such as Athena (Bernius et al., 2022), where the feedback actioning, such as acceptance or rejection, is easier for students because it is issued more granularly. Additionally, in our study students were typically required to perform more actions compared to these systems, which might also explain differences in the feedback action rate reported in the previous research. Lastly, our results should be seen in light of the broader literature on AI assistance interfaces. In particular, this literature suggests that depending on the nature of the resource editing activity, i.e., brainstorming ideas versus finalizing a resource, an efficient user interface that aids in both stages of the resource activity would offer diverse capabilities depending on the current stage (Laban et al., 2023). The relatively high actionability of students based on the AI assistance in our study could be explained by the fact that the assistance was issued to students once they had already established ideas and foundations for the resource creation and would have benefited more from granular suggestions in place of question body and option placeholders (Lee et al., 2024). This is as opposed to the benefit of the AI assistance issued in the sidebar at the beginning of the resource editing activities, which would have given students ideas on how to start their work on resource creation.

RQ3: How did students perceive the feedback provided by the AI? Our results indicate that 79% of students who responded rated the AI feedback to be helpful. A majority of comment segments (75%) said that they valued the feedback. In particular, students mentioned that they liked the concrete suggestions for improvement that the feedback provided, such as suggestions on writing better questions and options and improving writing quality. Our results are aligned with previous research suggesting that students value task-oriented actionable advice (Lim et al., 2020). However, numerous comment segments (25%) mentioned disagreements with the feedback or that it was inaccurate or unclear or contained redundant information. This is an authentic challenge when it comes to incorporating GenAI in educational technology, as was reported in other studies incorporating GenAI aiming to help students (Kazemitabaar et al., 2024). It is important to note that while we are relatively early in the proliferation of GenAI, most of the students likely had lots of exposure to numerous AI assistants and had already built up an intuition around its limitations. Nazaretsky and colleagues' (2024) results suggested that when students recognized that feedback was issued by AI, their perceptions of its quality decreased. In our study, students were well aware that the feedback provided to them was AI generated, which might be a confounding factor explaining the proportion of negative perceptions.

6.2 Implications and Lessons Learned

Implications for practice. Our study indicates the feasibility of equipping students with AI assistance during co-creation, which uses GenAI to deliver concise feedback incorporating detailed suggestions while preserving a positive tone. We hope that our study will pave the way for more fine-grained experimentation and authentic deployments of AI assistance similar to Jurenka and colleagues (2024). We suggest that designers and researchers who consider implementing AI support into their learning analytics and educational technologies differentiate between AI assistance support during the beginning of the resource editing activities, where ideation takes place, and at the final resource creation stage, where refinement happens. We would recommend supporting these two stages of resource editing using different user interfaces and different prompts if GenAI is involved. For instance, during the ideation stage, designers could deliver AI assistance in the sidebar, similar to the process in our study.

Since our results suggested that the majority of students tend to submit their post-feedback work without any edits, we see two strategies that could be used during the refinement stage of resource editing to improve the actionability rate. First, more granular interfaces and different suggestions might be required to ensure higher actionability and resource quality, as exemplified by writing assistants, where suggestions are directly overlaid on the text (Knight et al., 2020). As recent research suggests that students do not critically engage with GenAI results (Shibani et al., 2024), the second strategy could be to implement specific scaffolds to enable critical engagement with AI feedback, similar to the process in Shibani and colleagues (2022). These scaffolds could be implemented in the form of prompts asking students to structurally reflect on their agreement with the issued feedback. As such, these prompts could bridge the gap between students with a lack of substantial critical engagement skills, which include being able to question, analyze, and make informed decisions regarding the appropriateness of the feedback itself, as well as having intuition around GenAI limitations, which powers the feedback system.

In order to ensure that students consider and act when necessary based upon AI assistance delivered via feedback, designers and researchers could do the following. To overcome the challenge of students neglecting high-quality AI-issued feedback, Nazaretsky and colleagues (2024) suggested incorporating teachers in the feedback loop, where the initial feedback would be generated by AI, and teachers could provide oversight or make necessary changes. This is consistent with the vision of the instructor's role in the delivery of AI-generated feedback reported by Pozdniakov and colleagues (2024). Alternatively, to overcome under-reliance on high-quality AI-generated feedback, designers could make AI assistance more personalized by incorporating data from LA, especially since previous research emphasized the benefits of feedback personalization on feedback recipience and course outcomes (Pardo et al., 2019; Iraj et al., 2021).

Implications for research. Further research is needed to explore alternative strategies that could help ensure that students consider high-quality AI feedback effectively. To partially address the issue of students relying on AI-generated feedback that may contain hallucinations, the currently deployed RiPPLE shows a notification message before providing AI assistance: "The feedback is AI-generated and may contain inaccuracies. Use it with caution and apply your domain expertise to assess its accuracy." We acknowledge that this measure alone is not sufficient to fully protect students from following potentially misleading suggestions produced by LLM model hallucinations. In this light, we see three important future research directions. First, future research could investigate how students with varying levels of general feedback literacy and AI literacy (Carless & Boud, 2018; C. Wang et al., 2024) interact with inaccurate AI feedback. Second, future work could focus on exploring how AI feedback tools should scaffold students to enable them to critically engage with AI feedback (Shibani & Buckingham Shum, 2024). Third, future work could explore the design space related to learning tasks that explicitly encourage critical interaction with AI, rather than treating it as a tool for simple task completion (Denny, Leinonen, et al., 2024). It would also be beneficial to have a deeper understanding of students' experiences in the moment when using AI assistance. As such, future work could employ automatic detectors of educationally relevant affective states. These detectors would enable the collection of qualitative data on students' immediate experiences (Baker et al., 2021), offering valuable insights into designing more effective student–AI interactions. Future research could also identify factors that contribute to students avoiding AI assistance (C. Wang et al., 2024).

6.3 Ethical Considerations

We identify two primary ethical considerations related to the reliance on proprietary GenAI models: *privacy when using AI feedback* and *intellectual property risks* (Bommasani et al., 2022). Currently, enterprise-level GenAI models offer only limited secure access to institutional educational data. While locally deployed models could mitigate privacy risks, they often lag in performance compared to on-premise models and require custom integrations. In our study, no personal student data was included when requesting AI feedback, as demonstrated in the prompt templates provided in Appendix C, Listings 1 and 2. Thus, we do not perceive any privacy risks or violations. However, the content of student-created learning resources was used in the prompts, introducing a potential risk related to intellectual property. This implies that artifacts produced by students could be leveraged as training data for future iterations of GenAI models. We did not opt in to allow the data to be used for training purposes, as the AI assistance deployed in RiPPLE used an API (application programming interface) rather than a visual interface. Nevertheless, we acknowledge that this does not fully address the ethical concern of using student-generated content for the benefit of third parties without explicit student consent.

Two additional ethical concerns, namely, the *ability to opt out of AI assistance* and *equality of outcomes* (Yan et al., 2023), are relevant for both proprietary and locally deployed GenAI models. For the first concern, students in our study had the option to disable AI assistance and were made aware of this capability. Regarding equality, deploying imperfect GenAI tools raises questions about fairness and the consequences of their use in educational contexts. In our study, all students used the same predefined prompt, without the option to customize it, ensuring equal opportunities to receive feedback of consistent quality. This approach prevented students with higher AI literacy, such as superior prompting skills, from gaining an advantage by receiving better feedback.

6.4 Limitations

While our study provides valuable insights into the use of GenAI for providing feedback on student-created resources, several limitations need to be acknowledged. First, the results obtained in this study are based on specific prompts designed for creating and refining MCQs. It remains uncertain whether similar outcomes would be observed with different types of prompts or tasks, such as essay writing, project work, or problem-solving exercises. Second, the evaluation of AI feedback helpfulness and student perceptions relies on self-reported data from students, which is subject to biases and requires triangulation with other measures, such as students' skills at recognizing poor and good feedback, students' actual performance improvement, or engagement measures, providing a more comprehensive understanding. Third, the study was conducted within a specific educational context, involving higher education students at an Australian university. The findings might not be directly transferable to other educational settings, age groups, or contexts. By acknowledging these limitations, we hope to provide a foundation for future research to address these gaps and further enhance our understanding of the role of GenAI in educational feedback.

While the use of LLMs for classifying feedback elements has shown promising results, our study highlights several limitations that emphasize the necessity of human oversight to ensure classification accuracy. Although the LLMs we tested achieved moderate to high agreement across most categories, inconsistencies arose, particularly in complex classifications (R. Wang et al., 2023). Additionally, the lack of transparency in LLM decision-making processes makes it difficult to trace how specific classifications were determined. This stands in stark contrast to more transparent NLP methods, which, when combined with techniques like LIME, offer clearer insights into the reasoning behind algorithmic decisions (Hutt et al., 2024; Huang et al., 2025). These challenges underscore the current limitations of LLMs in handling nuanced and multifaceted feedback without human intervention. While recent research has demonstrated improved effectiveness by using pipelines with multiple LLMs (Gosmar & Dahl, 2025), the need to employ additional models to resolve disagreements further suggests that LLMs alone may not suffice for achieving optimal accuracy. Therefore, integrating human expertise to review and refine LLM-generated classifications is critical (Xu et al., 2024). This hybrid approach not only reduces the risk of misclassification but also leverages the complementary strengths of human judgment and LLM capabilities, ultimately enhancing the accuracy and reliability of feedback analysis.

7. Conclusion

This study explored aspects of AI assistance for students to support co-creation of educational content, how students acted upon receiving AI help, and how students perceived AI assistance. Our findings demonstrate that, first, AI assistance on both discipline-related knowledge and resource specifics is characterized by using simpler vocabulary in content summary and positive appraisals, while using more complex words in suggestions for improvement, and having a positive sentiment. Second, students would typically attempt to take action on the AI feedback in less than half (35%) of the resource editing sessions. Third, students generally found the AI assistance helpful, and they appreciated that feedback included suggestions to improve specific sections of the resource. Students who did not appreciate AI feedback did so mainly due to feedback being inaccurate and containing redundant information. These findings illustrate challenges encountered in integrating AI into the educational setting, including achieving consistent results, personalizing prompts for students, and developing intuitive user interfaces for human–AI collaboration.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research was at least in part funded by The Australian Research Council's Industrial Transformation Training Centre for Information Resilience (CIRES) (IC200100022)

References

Abdi, S., Khosravi, H., Sadiq, S., & Demartini, G. (2021). Evaluating the quality of learning resources: A learnersourcing approach. *IEEE Transactions on Learning Technologies*, *14*(1), 81–92. https://doi.org/10.1109/TLT.2021.3058644

Baker, R. S., Nasiar, N., Ocumpaugh, J. L., Hutt, S., Andres, J. M. A. L., Slater, S., Schofield, M., Moore, A., Paquette, L., Munshi, A., & Biswas, G. (2021). Affect-targeted interviews for understanding student frustration. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 52–63, Vol. 12748). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_5

Bernius, J. P., Krusche, S., & Bruegge, B. (2022). Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, *3*, 100081. https://doi.org/10.1016/j.caeai.2022.100081

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2022). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://doi.org/10.48550/arXiv.2108.07258

Bovill, C. (2020). Co-creation in learning and teaching: The case for a whole-class approach in higher education. *Higher Education*, *79*(6), 1023–1037. https://doi.org/10.1007/s10734-019-00453-w

Bovill, C., Cook-Sather, A., & Felten, P. (2011). Students as co-creators of teaching approaches, course design, and curricula: Implications for academic developers. *International Journal for Academic Development*, *16*(2), 133–145. https://doi.org/10.1080/1360144X.2011.568690

Buckingham Shum, S., Ferguson, R., & Martinez-Maldonaldo, R. (2019). Human-centred learning analytics. *Journal of Learning Analytics*, *6*(2), 1–9. https://doi.org/10.18608/jla.2019.62.1

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

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Chen, B., Zhu, X., & Díaz Del Castillo, F. (2023). Integrating generative AI in knowledge building. *Computers and Education: Artificial Intelligence*, *5*, 100184. https://doi.org/10.1016/j.caeai.2023.100184

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. (2023). What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures*, *17*(2), 111–130. https://doi.org/10.1080/19312458.2023.2167965

Choi, S., Lee, H., Lee, Y., & Kim, J. (2024). VIVID: Human-AI collaborative authoring of vicarious dialogues from lecture videos. *arXiv preprint arXiv:2403.09168*. http://arxiv.org/abs/2403.09168

Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *RadioGraphics*, *26*(2), 543–551. https://doi.org/10.1148/rg.262055145

Cook-Sather, A. (2014). Student-faculty partnership in explorations of pedagogical practice: A threshold concept in academic development. *International Journal for Academic Development*, *19*(3), 186–198. https://doi.org/10.1080/1360144X.2013.805694

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, *42*(3-4), 541–561. https://doi.org/10.1111/1467-9817.12283

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In M. Chang, N.-S. Chen, R. Kuo, G. Rudolph, D. G. Sampson, & A. Tlili (Eds.), *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies* (ICALT), 10–13 July 2023, Orem, Utah, USA (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Darvishi, A., Khosravi, H., Abdi, S., Sadiq, S., & Gašević, D. (2022). Incorporating training, self-monitoring and AI-assistance to improve peer feedback quality. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (L@S 2022), 1–3 June 2022, New York, New York, USA (pp. 35–47). ACM. https://doi.org/10.1145/3491140.3528265

Darvishi, A., Khosravi, H., & Sadiq, S. (2021). Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. *Proceedings of the Eighth ACM Conference on Learning @ Scale* (L@S 2021), 22–25 June 2021, online, 139–150. https://doi.org/10.1145/3430895.3460129

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, *210*, 104967. https://doi.org/10.1016/j.compedu.2023.104967

Denny, P., Gulwani, S., Heffernan, N. T., Käser, T., Moore, S., Rafferty, A. N., & Singla, A. (2024). Generative AI for education (GAIED): Advances, opportunities, and challenges. *arXiv preprint arXiv:2402.01580*. https://doi.org/10.48550/arXiv.2402.01580

Denny, P., Khosravi, H., Hellas, A., Leinonen, J., & Sarsa, S. (2023). Can we trust AI-generated educational content? Comparative analysis of human and AI-generated learning resources. *arXiv preprint arXiv:2306.10509*. https://doi.org/10.48550/arXiv.2306.10509

Denny, P., Leinonen, J., Prather, J., Luxton-Reilly, A., Amarouche, T., Becker, B. A., & Reeves, B. N. (2024). Prompt problems: A new programming exercise for the generative AI era. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education* (SIGCSE 2024), 20–23 March 2024, Portland, Oregon, USA (pp. 296–302, Vol. 1). ACM. https://doi.org/10.1145/3626252.3630909

Di Mitri, D., Schneider, J., & Drachsler, H. (2022). Keep me in the loop: Real-time feedback with multimodal data. *International Journal of Artificial Intelligence in Education*, *32*(4), 1093–1118. https://doi.org/10.1007/s40593-021-00281-z

Dollinger, M., Brazil, J., Matthews, K., & Khosravi, H. (2024). Is scalable "students as partners" possible?: Towards large-scale, inclusive, in-class partnerships. In T. Cochrane, V. Narayan, E. Bone, C. Deneen, M. Saligari, K. Tregloan, & R. Vanderburg (Eds.), *Proceedings of the 2024 ASCILITE Conference: Navigating the Terrain: Emerging Frontiers in Learning Spaces, Pedagogies, and Technologies* (ASCILITE 2024), 1–4 December 2024, Melbourne, Australia (pp. 1–11). ASCILITE Publications. https://doi.org/10.14742/apubs.2024.983

Dollinger, M., & Lodge, J. (2020). Student-staff co-creation in higher education: An evidence-informed model to support future design and implementation. *Journal of Higher Education Policy and Management*, *42*(5), 532–546. https://doi.org/10.1080/1360080X.2019.1663681

Dollinger, M., & Lodge, J. M. (2018). Co-creation strategies for learning analytics. In *Proceedings of the Eighth International Conference on Learning Analytics and Knowledge* (LAK 2018), 7–9 March 2018, Sydney, Australia (pp. 97–101). ACM. https://doi.org/10.1145/3170358.3170372

Doyle, E., Buckley, P., & Whelan, J. (2019). Assessment co-creation: An exploratory analysis of opportunities and challenges based on student and instructor perspectives. *Teaching in Higher Education*, *24*(6), 739–754. https://doi.org/10.1080/13562517.2018.1498077

Fan, Y., Jovanović, J., Saint, J., Jiang, Y., Wang, Q., & Gašević, D. (2022). Revealing the regulation of learning strategies of MOOC retakers: A learning analytic study. *Computers & Education*, *178*, 104404. https://doi.org/10.1016/j.compedu.2021.104404

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Gao, M., Hu, X., Ruan, J., Pu, X., & Wan, X. (2024). LLM-based NLG evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*. https://doi.org/10.48550/arXiv.2402.01383

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. https://doi.org/10.48550/arXiv.2312.10997

Gatta, R., Lenkowicz, J., Vallati, M., Rojas, E., Damiani, A., Sacchi, L., De Bari, B., Dagliati, A., Fernandez-Llatas, C., Montesi, M., Marchetti, A., Castellano, M., & Valentini, V. (2017). pMineR: An innovative R library for performing process mining in medicine. In A. ten Teije, C. Popow, J. Holmes, & L. Sacchi (Eds.), *Artificial intelligence in medicine. AIME 2017. Lecture notes in computer science* (pp. 351–355, Vol. 10259). Springer. https://doi.org/10.1007/978-3-319-59758-4_42

Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., & Drachsler, H. (2024). From the automated assessment of student essay content to highly informative feedback: A case study. *International Journal of Artificial Intelligence in Education*, *34*, 1378–1416. https://doi.org/10.1007/s40593-023-00387-6

Gosmar, D., & Dahl, D. A. (2025). Hallucination mitigation using agentic AI natural language-based frameworks. *arXiv preprint arXiv:2501.13946*. https://doi.org/10.48550/arXiv.2501.13946

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., Lee, T. Y., Hong, H., Kim, J., Ahn, S.-Y., & Oh, A. (2023). FABRIC: Automated scoring and feedback generation for essays. *arXiv preprint arXiv:2310.05191*. https://doi.org/10.48550/arXiv.2310.05191

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

Hilton, C. B., Goldwater, M. B., Hancock, D., Clemson, M., Huang, A., & Denyer, G. (2022). Scalable science education via online cooperative questioning. *CBE—Life Sciences Education*, *21*(1), ar4. https://doi.org/10.1187/cbe.19-11-0249

Huang, K., Mello, R. F., & Junior, C. P. (2025). That's what RoBERTa said: Explainable classification of peer feedback. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge* (LAK 2025), 3–7 March 2025, Dublin, Ireland (pp. 880–886). ACM. https://doi.org/10.1145/3706468.37065

Hutt, S., DePiro, A., Wang, J., Rhodes, S., Baker, R. S., Hieb, G., Sethuraman, S., Ocumpaugh, J., & Mills, C. (2024). Feedback on feedback: Comparing classic natural language processing and generative AI to evaluate peer feedback. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 55–65). ACM. https://doi.org/10.1145/3636555.3636850

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (ICWSM-14), 1–4 June 2014, Ann Arbor, Michigan, USA (pp. 216–225, Vol. 8). AAAI. https://doi.org/10.1609/icwsm.v8i1.14550

Hwang, K., Challagundla, S., Alomair, M. M., Chen, L. K., & Choa, F.-S. (2024). Towards AI-assisted multiple choice question generation and quality evaluation at scale: Aligning with Bloom's taxonomy. In *NeurIPS Workshop on Generative AI for Education (GAIED)* (NeurIPS 2023), 10–16 December 2023, New Orleans, Louisiana, USA. NeurIPS. https://gaied.org/neurips2023/files/17/17_paper.pdf

Iraj, H., Fudge, A., Khan, H., Faulkner, M., Pardo, A., & Kovanović, V. (2021). Narrowing the feedback gap: Examining student engagement with personalized and actionable feedback messages. *Journal of Learning Analytics*, *8*(3), 101–116. https://doi.org/10.18608/jla.2021.7184

Jurenka, I., Kunesch, M., McKee, K., Gillick, D., Zhu, S., Wiltberger, S., Phal, S. M., Hermann, K., Kasenberg, D., Bhoopchand, A., Anand, A., Pîslar, M., Chan, S., Wang, L., She, J., Mahmoudieh, P., Rysbek, A., Huber, A., Wiltshire, B., … Ibrahim, L. (2024). Towards responsible development of generative AI for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*. https://doi.org/10.48550/arXiv.2407.12687

Jury, B., Lorusso, A., Leinonen, J., Denny, P., & Luxton-Reilly, A. (2024). Evaluating LLM-generated worked examples in an introductory programming course. In N. Herbert & C. Seton (Eds.), *Proceedings of the 26th Australasian Computing Education Conference* (ACE 2024), 29 January–2 February 2024, Sydney, Australia (pp. 77–86). ACM. https://doi.org/10.1145/3636243.3636252

Kazemitabaar, M., Ye, R., Wang, X., Henley, A. Z., Denny, P., Craig, M., & Grossman, T. (2024). CodeAid: Evaluating a classroom deployment of an LLM-based programming assistant that balances student and educator needs. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, ePhoebe Toups Dugas, & I. Shklovski (Eds.), *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI 2024), 11–16 May 2024, Honolulu, Hawaii, USA (pp. 1–20). ACM. https://doi.org/10.1145/3613904.3642773

Keuning, H., Jeuring, J., & Heeren, B. (2019). A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education*, *19*(1). https://doi.org/10.1145/3231711

Khosravi, H., Denny, P., Moore, S., & Stamper, J. (2023). Learnersourcing in the age of AI: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*, *5*, 100151. https://doi.org/10.1016/j.caeai.2023.100151

Khosravi, H., Kitto, K., & Williams, J. J. (2019). RiPPLE: A crowdsourced adaptive platform for recommendation of learning activities. *Journal of Learning Analytics*, *6*(3), 91–105. https://doi.org/10.18608/jla.2019.63.12

Kim, J. (2015). *Learnersourcing: Improving learning with collective learner activity* [Doctoral dissertation, Massachusetts Institute of Technology].

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., Wight, R., Lucas, C., Sándor, Á., Kitto, K., Liu, M., Mogarkar, R. V., & Buckingham Shum, S. (2020). AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research*, *12*(1). https://doi.org/10.17239/jowr-2020.12.01.06

Laban, P., Vig, J., Hearst, M. A., Xiong, C., & Wu, C.-S. (2023). Beyond the chat: Executable and verifiable text-editing with LLMs. *arXiv preprint arXiv:2309.15337*. https://doi.org/10.48550/arXiv.2309.15337

Lahza, H., Khosravi, H., Demartini, G., & Gasevic, D. (2022). Effects of technological interventions for self-regulation: A control experiment in learnersourcing. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 542–548). ACM. https://doi.org/10.1145/3506860.3506911

Lee, U., Han, A., Lee, J., Lee, E., Kim, J., Kim, H., & Lim, C. (2024). Prompt Aloud!: Incorporating image-generative AI into STEAM class with learning analytics using prompt data. *Education and Information Technologies*, *29*, 9575–9605. https://doi.org/10.1007/s10639-023-12150-4

Leiker, D., Gyllen, A. R., Eldesouky, I., & Cukurova, M. (2023). Generative AI for learning: Investigating the potential of learning videos with synthetic virtual instructors [Series Title: Communications in Computer and Information Science]. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education. posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky. AIED 2023. Communications in computer and information science* (pp. 523–529, Vol. 1831). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36336-8_81

Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2024). The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*. https://doi.org/10.48550/arXiv.2401.03205

Liang, Y., & Matthews, K. E. (2023). Are Confucian educational values a barrier to engaging students as partners in Chinese universities? *Higher Education Research & Development*, *42*(6), 1453–1466. https://doi.org/10.1080/07294360.2022.2138276

Lim, L.-A., Dawson, S., Gašević, D., Joksimović, S., Pardo, A., Fudge, A., & Gentili, S. (2020). Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: An exploratory study of four courses. *Assessment & Evaluation in Higher Education*, *46*(3), 339–359. https://doi.org/10.1080/02602938.2020.1782831

Liu, F., & Lee, J. S. (2023). Hybrid models for sentence readability assessment. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA 2023), 13 July 2023, Toronto, Ontario, Canada (pp. 448–454). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.37

Lubicz-Nawrocka, T., & Bovill, C. (2023). Do students experience transformation through co-creating curriculum in higher education? *Teaching in Higher Education*, *28*(7), 1744–1760. https://doi.org/10.1080/13562517.2021.1928060

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, *47*(1), 141–179. https://doi.org/10.1162/coli_a_00398

Martinez-Maldonado, R., Echeverria, V., Schulte, J., Shibani, A., Mangaroska, K., & Buckingham Shum, S. (2020). Moodoo: Indoor positioning analytics for characterising classroom teaching. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education. AIED 2020. Lecture notes in computer science* (pp. 360–373, Vol. 12163). Springer. https://doi.org/10.1007/978-3-030-52237-7_29

Matcha, W., Uzir, N. A., Gasevic, D., & Pardo, A. (2020). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, *13*(2), 226–245. https://doi.org/10.1109/TLT.2019.2916802

Misiejuk, K., Kaliisa, R., & Scianna, J. (2024). Augmenting assessment with AI coding of online student discourse: A question of reliability. *Computers and Education: Artificial Intelligence*, *6*, 100216. https://doi.org/10.1016/j.caeai.2024.100216

Misiejuk, K., Wasson, B., & Egelandsdal, K. (2021). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior*, *117*, 106658. https://doi.org/10.1016/j.chb.2020.106658

Mohammadi, M., Tajik, E., Martinez-Maldonado, R., Sadiq, S., Tomaszewski, W., & Khosravi, H. (2024). Artificial intelligence in multimodal learning analytics: A systematic literature review. *ResearchGate preprint*. https://doi.org/10.13140/RG.2.2.16241.29281

Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *arXiv preprint arXiv:2306.10052*. https://doi.org/10.48550/arXiv.2306.10052

Moore, S., Nguyen, H. A., Chen, T., & Stamper, J. (2023). Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods [Series Title: Lecture Notes in Computer Science]. In O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, & T. Papathoma (Eds.), *Responsive and sustainable educational futures. EC-TEL 2023. Lecture notes in computer science* (pp. 229–245, Vol. 14200). Springer. https://doi.org/10.1007/978-3-031-42682-7_16

Moore, S., Nguyen, H. A., & Stamper, J. (2021, June 8). Examining the effects of student participation and performance on the quality of learnersourcing multiple-choice questions. In *Proceedings of the Eighth ACM Conference on Learning @ Scale* (L@S 2021), 22–25 June 2021, online (pp. 209–220). ACM. https://doi.org/10.1145/3430895.3460140

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. Ruipérez Valiente (Eds.), *Technology enhanced learning for inclusive and equitable quality education. EC-TEL 2024. Lecture notes in computer science* (Vol. 15159). Springer. https://doi.org/10.1007/978-3-031-72315-5_20

Nguyen, H., Stec, H., Hou, X., Di, S., & McLaren, B. (2023). Evaluating ChatGPT's decimal skills and feedback generation in a digital learning game. In O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, & T. Papathoma (Eds.), *Responsive and sustainable educational futures. EC-TEL 2023. Lecture notes in computer science* (pp. 278–293, Vol. 14200). Springer Nature. https://doi.org/10.1007/978-3-031-42682-7_19

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

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*. https://doi.org/10.48550/arXiv.2404.13076

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

Pozdniakov, S., Brazil, J., Abdi, S., Bakharia, A., Sadiq, S., Gasevic, D., Denny, P., & Khosravi, H. (2024). Large language models meet user interfaces: The case of provisioning feedback. *Computers and Education: Artificial Intelligence*, *7*, 100289. https://doi.org/10.1016/j.caeai.2024.100289

Qian, Y., & Lehman, J. (2017). Students' misconceptions and other difficulties in introductory programming: A literature review. *ACM Transactions on Computing Education (TOCE)*, *18*(1). https://doi.org/10.1145/3077618

Saldaña, J. (2015). *The coding manual for qualitative researchers*. SAGE.

Sanders, E. B.-N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *CoDesign*, *4*(1), 5–18. https://doi.org/10.1080/15710880701875068

Sarmiento, J. P., & Wise, A. F. (2022). Participatory and co-design of learning analytics: An initial review of the literature. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 535–541). ACM. https://doi.org/10.1145/3506860.350691

Shibani, A., & Buckingham Shum, S. (2024). AI-assisted writing in education: Ecosystem risks and mitigations. *arXiv preprint arXiv:2404.10281*. https://doi.org/10.48550/arXiv.2404.10281

Shibani, A., Knight, S., & Buckingham Shum, S. (2022). Questioning learning analytics? Cultivating critical engagement as student automated feedback literacy. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 326–335). ACM. https://doi.org/10.1145/3506860.3506912

Shibani, A., Knight, S., Kitto, K., Karunanayake, A., & Buckingham Shum, S. (2024). Untangling critical interaction with AI in students' written assessment. In F. F. Mueller, P. Kyburz, J. R. Williamson, & C. Sas (Eds.), *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (CHI EA 2024), 11–16 May 2024, Honolulu, Hawaii, USA (pp. 1–6). ACM. https://doi.org/10.1145/3613905.3651083

Singh, A., Brooks, C., & Doroudi, S. (2022). Learnersourcing in theory and practice: Synthesizing the literature and charting the future. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (L@S 2022), 1–3 June 2022, New York, New York, USA (pp. 234–245). ACM. https://doi.org/10.1145/3491140.3528277

Singh, A., Brooks, C., Wang, X., Li, W., Kim, J., & Wilson, D. (2024). Bridging learnersourcing and AI: Exploring the dynamics of student-AI collaborative feedback generation. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 742–748). ACM. https://doi.org/10.1145/3636555.3636853

Wang, C., Wang, H., Li, Y., Dai, J., Gu, X., & Yu, T. (2024). Factors influencing university students' behavioral intention to use generative artificial intelligence: Integrating the theory of planned behavior and AI literacy. *International Journal of Human–Computer Interaction*, 1–23. https://doi.org/10.1080/10447318.2024.2383033

Wang, R., Wirawarn, P., Goodman, N., & Demszky, D. (2023). SIGHT: A large annotated dataset on student insights gathered from higher education transcripts. *arXiv preprint arXiv:2306.09343*. https://doi.org/10.48550/arXiv.2306.09343

Xu, S., Huang, X., Lo, C. K., Chen, G., & Jong, M. S.-y. (2024). Evaluating the performance of ChatGPT and GPT-4o in coding classroom discourse data: A study of synchronous online mathematics instruction. *Computers and Education: Artificial Intelligence*, *7*, 100325. https://doi.org/10.1016/j.caeai.2024.100325

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*. https://doi.org/10.48550/arXiv.2309.06794

Zambrano, A. F., Liu, X., Barany, A., Baker, R. S., Kim, J., & Nasiar, N. (2023). From nCoder to ChatGPT: From automated coding to refining human coding. In G. Arastoopour Irgens & S. Knight (Eds.), *Advances in quantitative ethnography.*

ICQE 2023. Communications in computer and information science (pp. 470–485, Vol. 1895). Springer. https://doi.org/10.1007/978-3-031-47014-1_2

Appendices

Appendix A: RQ1

This appendix includes the results of topic modeling. We applied topic modeling separately for positive appraisals and improvement suggestions, resulting in two models. We experimented with 20, 15, and 10 topics, choosing 10 for both models to maintain content-agnostic, i.e., not focusing explicitly on the course subject, yet prominent feedback characteristics. Distribution-based outlier reduction attributed all but 5 appraisals to one of the 10 topics. Table 4 includes results of topic modeling applied to the corpus of positive appraisals ($N_{topic\_number} = 10$). Table 5 includes results of topic modeling for suggestions for improvements. We manually grouped semantically similar topics into three categories: "1-Clarity&Writing", "2-MCQ Design", and "3-Disciplinary Knowledge". Table 3 includes a representation of two merged topic models ($N_{topic\_number} = 3$), one for positive appraisals and one for suggestions for improvements.

Table 3. Overview of grouped topic modeling into three thematic categories.

Thematic category	Description	Positive appraisals examples	#	Suggestions examples	#
1-Clarity and Writing	Feedback resulting from evaluation of the linguistic correctness of the resource. It praises or makes suggestions whether the sentences are grammatically correct, the language is clear, and the question or explanation is well-written.	Document 1: The question body is clear and concise. Document 2: The resource uses an encouraging tone in the explanations. Document 3: The question body is clear and grammatically correct	1186	Document 1: Perhaps consider breaking down the definition into smaller parts for better clarity and easier absorption of information. Document 2: The explanations could be more concise and structured. Instead of lengthy paragraphs, consider breaking down the explanations into shorter, more digestible points. This enhances clarity and makes it easier for students to follow the reasoning behind each option Document 3: The question could be more meaningful without having to read all the options first.	1687
2-MCQ Design	Feedback resulting from evaluation of the quality of the multiple-choice questions. It praises or makes suggestions whether the questions are clear, the options are well-structured, and the explanations provide clear distinctions between correct and incorrect answers (distractors).	Document 1: The options are plausible and from the same category Document 2: All options marked as incorrect are indeed inaccurate , Document 3: The steps provided for finding the normal form are easy to follow and understand.	1316	Document 1: Explanation 3 should clearly explain why the option is incorrect. For example, Incorrect. The option includes the attribute 'orderNum' as underlined, but the SQL statement does not mention underlining any attribute. Document 2: 5 & 57 & 2-Revise incorrect options such that they are more plausible & Document 1: The options could be improved by ensuring they are more distinct. For instance, option 2 could be made more challenging by having a closer value margin between marginal cost and average total cost, making the decision less obvious. Document 3:	667
3-Disciplinary content	Feedback resulting from evaluation of the accuracy and depth of the subject matter content. It praises or makes suggestions if the explanations are accurate, easy to understand, and provide a clear understanding of the concept being tested or explained. It also assesses whether the resource covers the necessary disciplinary content in detail.	Document 1: The options include common ways to increase statistical power , Document 2: The suggestion for creating a structured framework to avoid entrepreneurial pitfalls is a positive contribution. Document 3: Provides a clear distinction between simply knowing a language for passing exams versus actually using it in real-life situations.	547	Document 1: Exploring how external factors like market conditions or industry trends may influence the effectiveness of scaling up versus scaling deep could provide a more holistic understanding. Document 2: It may be helpful to delve deeper into the competitive landscape and how Polarity plans to differentiate itself from potential competitors. Document 3: Document 2: The options could be enhanced by introducing more distractors that reflect common errors made in GDP calculations. For instance, including an option that miscalculates the nominal GDP by using the 2020 price instead of the 2021 price would provide a valuable misconception to address.	963

Table 4. Description of topic modeling over positive appraisals.

Topic #	# of Items	Name	Representative items	Top Terms
1	896	1-Clear writing and effective questioning in specific concepts	Document 1: The options cover different dependencies, allowing for a comprehensive assessment of knowledge. Document 2: Provides a detailed overview of PolarityTE's focus on regenerative medicine and innovative skin regeneration technologies. Document 3: The impact and challenges section provides a realistic view of the company's position in the market.	['specific' 'question clearly' 'provides' 'concept' 'focuses' 'provides clear' 'accurate provides' 'clearly written' 'written' 'addresses']
4	110	1-Clear, concise and grammatically correct question	Document 1: The question body is clear and concise. Document 2: The question body is clear and grammatically correct. Document 3: The question body is clear and grammatically correct.	['correct question' 'written grammatically' 'grammatically correct' 'grammatically' 'marked correct' 'body clearly' 'question body' 'marked' 'body' 'written']
5	96	1-Encouraging and clear explanations	Document 1: There is an encouraging tone in the explanations. Document 2: The resource uses an encouraging tone in the explanations. Document 3: The resource uses an encouraging tone in the explanations.	['tone' 'encouraging' 'use' 'resource uses' 'uses' 'provide accurate' 'accurate information' 'explanations easy' 'information' 'explanations']
8	84	1-Framing questions clearly	Document 1: The resource is framed as a question rather than an incomplete statement. Document 2: The resource is framed as a question rather than an incomplete statement. Document 3: The resource is framed as a question rather than an incomplete statement.	['question question' 'statement' 'question clearly' 'clearly written' 'written' 'clearly' 'clear meaningful' 'resource' 'question' 'clear scenario']
0	1164	2-Clear and accurate MCQ option explanations	Document 1: The steps provided for finding the normal form are easy to follow and understand. Document 2: The resource highlights the potential of SkinTE to challenge existing treatments, introduce innovative solutions, and capitalize on market opportunities. Document 3: The examples provided, such as PolarityTE's SkinTE® product and their expansion into new markets like diabetic foot ulcers, offer practical illustrations of the concepts discussed.	['easy understand' 'explanations accurate' 'easy' 'understand' 'accurate' 'accurate provide' 'explanations provided' 'option accurate' 'provided' 'option']
6	116	2-Plausible and incorrect options	Document 1: The options are plausible and from the same category. Document 2: The options are plausible and from the same category. Document 3: The options are plausible and from the same category.	['options plausible' 'plausible' 'incorrect options' 'use common' 'incorrect' 'challenging' 'educational' 'use' 'question options' 'options']
7	36	2-Correct identification of incorrect MCQ options	Document 1: All options marked as incorrect are indeed inaccurate. Document 2: All options marked as incorrect are indeed inaccurate. Document 3: All options marked as incorrect are indeed inaccurate.	['marked incorrect' 'options marked' 'marked' 'inaccurate' 'marked correct' 'incorrect' 'options' 'incorrect explanations' 'factually inaccurate' 'factually']
2	225	3-Use of common mistakes and misconceptions in options creation	Document 1: The options include common aspects of emotional intelligence. Document 2: The options include common genetic markers associated with different conditions. Document 3: The options include common ways to increase statistical power.	['mistakes misconceptions' 'common mistakes' 'mistakes' 'include common' 'options include' 'include' 'options use' 'common' 'use common' 'misconceptions']
3	286	3-Clear question formulation using disciplinary knowledge	Document 1: It highlights Apple's strengths in creating a seamless user experience, building a strong ecosystem, and prioritizing privacy and user trust. Document 2: The suggestion for creating a structured framework to avoid entrepreneurial pitfalls is a positive contribution. Document 3: Provides a clear distinction between simply knowing a language for passing exams versus actually using it in real-life situations.	['presents clear' 'resource presents' 'clear question' 'presents' 'resource' 'correct question' 'creating' 'clear grammatically' 'question options' 'clear']
9	36	3-Incorrect MCQ options are factually incorrect	Document 1: The resource includes options that are factually incorrect. Document 2: The incorrect options are factually incorrect. Document 3: The incorrect options are factually incorrect.	['options inaccurate' 'options factually' 'inaccurate' 'factually incorrect' 'factually inaccurate' 'factually' 'incorrect options' 'incorrect' 'incorrect explanations' 'resource includes']

Table 5. Description of topic modeling over suggestions.

Topic #	# of Items	Name	Representative items	Top Terms
0	1605	1-Writing Clear and Concise Question Bodies	Document 1: The question body could be more concise by avoiding unnecessary details about the specific schema elements. For instance, you could simplify the schema representation to focus solely on the functional dependencies being tested. Document 2: The question body could be more concise. Instead of asking Is this good for an economy? consider asking What does a Gini Coefficient of 0.67 indicate about income equality in South Africa? This revision could make the question more direct and focused. Document 3: The question could be more meaningful without having to read all the options first. For example, rephrasing the question to focus on identifying cellular mediators of inflammation directly could enhance clarity.	['details' 'body' 'clearly' 'question body' 'help' 'example instead' 'help respondent' 'concise' 'respondent' 'understand']
6	82	1-Improve readibility of resource sections	Document 1: Perhaps consider breaking down the definition into smaller parts for better clarity and easier absorption of information. Document 2: The explanations could be more concise and structured. Instead of lengthy paragraphs, consider breaking down the explanations into shorter, more digestible points. This enhances clarity and makes it easier for students to follow the reasoning behind each option. Document 3: Consider breaking down the content into smaller sections for easier digestion.	['readability' 'complex' 'sections' 'points' 'content' 'comprehension' 'easier' 'benefit' 'enhance readability' 'better']
3	272	2-Make improvements for correct/incorrect options	Document 1: Explanation 3 should clearly explain why the option is incorrect. For example, Incorrect. The option includes the attribute 'orderNum' as underlined, but the SQL statement does not mention underlining any attribute. Document 2: The explanations should provide further details on why the incorrect options are incorrect. For example, explain that urinary retention is a symptom of Cauda equina syndrome, but it is not a red flag symptom that requires immediate medical attention. Document 3: The explanations for the incorrect options should clearly explain why the option is incorrect. For example, instead of simply stating that a certain attribute is not the primary key, provide a more detailed explanation of why it cannot be part of a superkey set. This will help clarify the concept for the student and provide a more comprehensive understanding.	['incorrect option' 'option incorrect' 'attribute' 'relationship' 'incorrect example' 'represents' 'represented' 'explain' 'plausible category' 'category']
5	318	2-Revise incorrect options such that they are more plausible	Document 1: The options could be improved by ensuring they are more distinct. For instance, option 2 could be made more challenging by having a closer value margin between marginal cost and average total cost, making the decision less obvious. Document 2: Option 1 could be improved by making it more plausible. A significant increase in household income would typically not lead to a family facing a housing crisis. Consider revising this option to align more closely with the theme of the question. Document 3: Option 5 could be improved by making it more plausible. For instance, instead of stating that two different entities cannot have attributes with the same name (which is false), a better distractor could be created by stating that two different entities cannot have the same key attribute. This adjustment would make the option more challenging and align it with common misconceptions in ER diagram design.	['example option' 'improved' 'revised' 'instance option' 'option improved' 'improved making' 'misconception related' 'making' 'plausible' 'misconception']
8	77	2-Make changes to correct answer	Document 1: The selected answer is not factually correct. The correct proportion is 0.37, not 0.36. The explanation should mention that the answer was rounded off to two decimal places as instructed in the question. Document 2: The selected answer is factually accurate and indeed the best answer among the options provided. Document 3: The correct option could be evaluated to determine if it is the best answer out of all the options presented. Based on my evaluation, option 2 is accurate and the best answer.	['answer options' 'best answer' 'best' 'answer' 'marked' 'marked correct' 'options presented' 'correct correct' 'accurate' 'answer factually']
7	55	3-Writing clearer explanations for options	Document 1: In the CPI formula, consider explicitly defining P and Q to avoid confusion for learners who may not be familiar with these symbols. Document 2: The explanation for Option 4 could be expanded to provide a clearer reason why it is incorrect. For example, Fatty acyl-carnitine is actually a key step in transporting long-chain fatty acids into the mitochondria, not a substance that restricts their entry." Document 3: Provide a step-by-step breakdown of how the limit of x^k equals 2 and 0 respectively, to connect the mathematical concepts with the conclusion that set A is not closed.	['equation' 'ratio' 'cardinality' 'step' 'explanation option' 'group' 'option factually' 'production' 'relates' 'fatty']
4	187	3-Add a brief explanation and examples for an option	Document 1: Provide more context or explanation about the significance of prompt engineering in the field of Language Models. Document 2: In the explanation section, you could provide a brief rationale or economic interpretation behind the concept of national savings to deepen understanding. Document 3: Consider providing a brief comparison with Universal Grammar and Behaviorist-based accounts to highlight the unique advantages of a functional approach more clearly.	['brief' 'section' 'brief explanation' 'consider adding' 'adding' 'faced' 'providing brief' 'challenges' 'consider providing' 'limitations']
1	509	3-Exemplify or illustrate with speicfic examples	Document 1: Exploring how external factors like market conditions or industry trends may influence the effectiveness of scaling up versus scaling deep could provide a more holistic understanding. Document 2: Consider providing examples or visual aids to further illustrate how input encoding, transformer layers, and attention mechanism work in practice. Document 3: Add examples to illustrate the concepts of fine-tuning and prompt engineering.	['examples' 'studies' 'illustrate' 'examples case' 'case studies' 'specific examples' 'case' 'providing specific' 'length' 'consider providing']
2	158	3-Add specific content with references to the resource	Document 1: It may be helpful to include more information on the regulatory landscape and any potential barriers to entry that Polarity could encounter in different markets for a more comprehensive analysis. Document 2: It may be helpful to delve deeper into the competitive landscape and how Polarity plans to differentiate itself from potential competitors. Document 3: It may be helpful to include comparative analyses with other biotech companies or disruptive innovators to enhance the depth of the discussion.	['helpful' 'include' 'beneficial' 'language' 'include brief' 'strategies' 'references' 'studies' 'language acquisition' 'acquisition']
9	54	3-Content-specific suggestions for different parts of MCQs	Document 1: Check the final decimal results for growth rate and GDP deflator calculations to ensure precision. Document 2: The options could be enhanced by introducing more distractors that reflect common errors made in GDP calculations. For instance, including an option that miscalculates the nominal GDP by using the 2020 price instead of the 2021 price would provide a valuable misconception to address. Document 3: The explanation for each option could be expanded to include why the specific values were chosen. For instance, explaining why the nominal GDP is calculated using the 2021 price rather than the 2020 price would deepen the understanding of the calculation process.	['gdp' 'calculation' 'real' 'calculations' 'price' 'calculated' 'growth' 'calculation process' 'explain incorrect' 'components']

Appendix B: RQ3

Table 6 includes the coding scheme used for the analysis of students' comments. It includes a list of themes, codes, descriptions, and examples derived from the data. Each code is defined and exemplified to ensure clarity and consistency in the interpretation of qualitative results.

Table 6. Coding scheme used for coding of students' comments.

Theme	Code name	Description	Example
Valuable aspects of AI feedback	1-General valuable aspects	Comments refers to the feedback being valuable without clear justification or categorization for why it was valuable.	The feedback is very helpful and enlightening to me.
	2-Specific valuable aspects	Comments refer to the feedback being valuable with a clear justification or categorization for why it was valuable.	
	2.1-Useful suggestion(s) provided	Comment refers to suggestions for improvement in the feedback being valuable or the feedback leading to some general change in the students' work.	Suggestions are very useful.
	2.2-Led to improvements	Comment refers to the feedback making the student (a) aware that a specific aspect of their work requires changes and/or (b) changes a specific part of their resource.	
	2.2.1-Answer and distractor(s)	Comment refers to awareness or action of changes required to the answer and/or distractors.	This real feedback was really good and helped me understand that my interpretation of the answer might not be quite right.
	2.2.2-Explanations	Comment refers to awareness or action of changes required to the explanations.	The suggestion for expanding on the explanation for the correct answer was helpful
	2.2.3-Question body	Comment refers to awareness or action of changes required to the question body or stem.	The feedback helped me improve the body of my question.
	2.2.4-Writing quality	Comment refers to awareness or action of changes required to writing quality.	The AI feedback was helpful and beneficial towards improving my question to ensure it was articulated correctly and that others would be able to understand it when answering.
	2.2.5-Other	Comment refers to awareness or action of changes required to various other aspects that could not be categorized further.	The feedback helped me improve the difficulty of the question.
Issues and challenges of AI feedback	3-General issues and challenges	Comment refers to issues or challenges with the feedback without any specific justification.	It was ok feedback
	4-Specific issues and challenges	Comment refers to issues or challenges with the feedback with a specific justification.	
	4.1-Disagree with feedback	Comment refers to a disagreement with the feedback provided.	I don't know if all of the improvements would actually make it better.
	4.2-Inaccurate	Comment refers to incorrect information provided.	The last point mentioned was incorrect.
	4.3-Misunderstanding of author aim or task	Comment refers to feedback that did not match the student's purpose of the question or task requirements.	The AI feedback misses the point of the question. This is a simple remembering question - the AI feedback is geared towards a more complex question and is not at all applicable.
	4.4-Redundant	Comment refers to feedback that was already implemented by the student or had duplicate information.	Some of the feedback that was given I had already done in my question creation.
	4.5-Unclear	Comment refers to feedback that was uninterpretable or unactionable.	It is difficult to comprehend what it meant by simplifying the language.
	4.6-Other	Comment refers to issues or challenges with the feedback that could not be categorized further.	The second point is impractical

Appendix C: Prompt used

This appendix details the prompt used for implementing AI assistance in our study. Following the framework for pedagogical incorporation of GenAI described in the context section, we aimed to support students during the early stages of co-creation by addressing knowledge and skills gaps. The prompt was designed based on feedback design principles and explicit criteria for generating high-quality outputs. We utilized OpenAI's gpt-3.5 turbo model with a few-shot prompting technique, ensuring the feedback was kind, constructive, specific, and actionable. The final prompt, including all major elements, is presented in Listing 1, containing the main prompt body, and Listing 2, containing few-shot examples incorporated into the prompt.

Listing 1. Few-shot Prompt Used to Generate Feedback

Step 1 – Main Prompt Body

```
You are an expert exam question writer and tutor. Your student has been tasked with creating a
multiple-choice question. First, work out your own solution to the multiple-choice question asked.
Then compare your solution to the option the student said was the answer and evaluate if the option
the student said was correct is the one best answer out of all the options presented. Don't decide
if the option the student marked as the correct option is actually the most correct answer until you
have done the question yourself. If you evaluate the student has selected the correct answer,
mention this in the positives section. Remember, a multiple-choice question should have one or more
options or distractors that are factually incorrect. Do not comment if a distractor or option marked
as incorrect is factually incorrect. Your job is to then provide feedback to the student to help
improve their question using the criteria below. You must be indirect and unsure about your
considerations; use tentative statements like "could" "may" "perhaps" "consider". Your feedback must
be kind, constructive, specific and very actionable. For each suggestion, you must provide only one
example of how the suggestion can be applied with direct reference to the input provided. Write
question instead of "stem". Write options instead of "distractors". You must mention if the answer
is correct and the one best answer OR if it was incorrectly selected. You must check if the
distractor is truly inaccurate or not the best answer.

You must provide examples for HOW to apply your suggestions based on the users input. For example,
the question body could be more clearly written by avoiding unnecessary details. For example,
instead of mentioning the specific colors of the levers, the question could focus on the general
concept of using punishment to encourage a desired behavior. The explanation for the correct answer
should provide further details on why the option is correct. For example, "No, punishment alone does
not provide information on what behavior is desired. Consider rephrasing the question to make it
more concise and meaningful without having to read all the options first. For example, "Will
punishment alone create the desired association in the experiment?"

You must comment on the quality of the question body, options, and explanations using the following
criteria:

Criterion 1: Quality of question (or stem)
The question is clearly written and grammatically correct.
The question should be meaningful without having to read all the options first.
The question is not negatively worded or includes double negatives.

Criterion 2: Quality of options (or distractors)
The options use common mistakes/misconceptions.
The options are plausible and from the same category.
The options do not include 'None of the above'.
The options marked as correct are the best answers.
Options not marked as correct are factually inaccurate.
Options should avoid clues that give away the correct answer, such as grammatical cues, word
repetition, absolute terms or negative wording. If the option is not marked as correct, the
explanation should outline why the option is incorrect.

Criterion 3: Quality of explanation
The explanations are accurate.
The explanations are easy to understand.
The explanations use an encouraging tone.
The explanations clearly explain why the option is correct or incorrect.

Use the template and examples below to structure your response but DO NOT copy them verbatim or too
closely:

<p>The aim of this resource is to assess x.</p>

<h1><strong><span style="font-size: 12pt;"> Positives </span></strong></h1>
```

```
    <ul>
      <li>Positive aspect and why this has learning benefits.</li>
      <li>Positive aspect and why this has learning benefits.</li>
      <li>Positive aspect and why this has learning benefits</li>
    </ul>

    <p><strong><span style="font-size: 12pt;"> Considerations </span></strong></p>
    <ul>
      <li>Specific, actionable, tentative suggestion. Explanation.</li>
      <li>Specific, actionable, tentative suggestion. Explanation.</li>
      <li>Specific, actionable, tentative suggestion. Explanation.</li>
    </ul>
```

Listing 2. Few-shot Prompt Used to Generate Feedback

Step 2 – Few-shot Examples Included in the Prompt

```
Example response 1:

    <h1><strong><span style="font-size: 12pt;">Summary/span></strong></h1>
    <p>The aim of this resource is to assess knowledge about the capital cities of [Country] and can
    contribute to student learning.</p>
    <h1><strong><span style="font-size: 12pt;"> Positives </span></strong></h1>
    <ul>
    <li>The resource is framed as a question rather than an incomplete statement. This improves the ease
      of answering the question.</li>
    <li>All options marked as incorrect are indeed inaccurate, helping accurately test the students'
    knoweldge.</li>
    <li>The explanations are accurate and easy to understand, which is uesful for struggling students to
      learn more.</li>
    </ul>
    <p><strong><span style="font-size: 12pt;"> Suggestions for improvement</span></strong></p>
    <ul>
    <li>The question body should be more clearly written and avoid any irrelevant details. This will
    help the respondent understand the questions more clearly. </li>
    <li>The incorrect options should be revised to ensure they are plausible and
    from the same category. For example, use another [Country] city, such as
    [city name], instead of including Paris. This is because unrealistic or
    humorous options increase the respondent's chance of guessing the
    answer.</li>

    <li>Explaining the incorrect options should provide further details on why the option is incorrect.
    For example, Incorrect; Sydney is one of the most populated cities in [Country] but was not selected
      as the capital city.&rdquo; Providing explanations on all options facilitates learning.</li>
    </ul>

Example response 2:

    <h1><strong><span style="font-size: 12pt;">Summary</span></strong></h1>
    <p>The aim of this resource is to assess knowledge about the brain region involved with processing
    emotion.</p>
    <h1><strong><span style="font-size: 12pt;"> Positives </span></strong></h1>
    <ul>
    <li>The resource present a clear question which facilitates learning.</li>
    <li>The options include common mistakes or misconceptions. This enhances the learning process by
    explicitly addressing and correcting misunderstandings, thereby improving overall test performance
    and confidenc </li>
    </ul>
    <p><strong><span style="font-size: 12pt;"> Suggestions for improvement</span></strong></p>
    <ul>
    <li>The question body should not be negatively worded. For example, instead of asking the respondent
      to select an option that is not true, ask the respondent to select the correct option. This is
    because using negatives can confuse students and lead to sentences that are difficult to interpret.
    If negatives can't be avoided, highlight the negative in the question's stem (e.g., in bold).</li>
    <li>Adjust the wording of all options to be a similar length to avoid giving away any clues to the
    answer. The correct option is longer than all other options, which might help respondents guess the
    answer.</li>
    <li>The selected answer is not factually correct. The answer should be the amygdala, as it is
    responsible for processing emotional responses.</li>
```