

Assessing the Proficiency of Large Language Models in Automatic Feedback Generation: An Evaluation Study

Wei Dai^a, Yi-Shan Tsai^a, Jionghao Lin^b, Ahmad Aldino^a, Hua Jin^a, Tongguang Li^a, Dragan Gašević^a, Guanliang Chen^a

^aMonash University, Melbourne, Australia

^bCarnegie Mellon University, USA

Abstract

Assessment feedback is important to student learning. Learning analytics (LA) powered by artificial intelligence exhibits profound potential in helping instructors with the laborious provision of feedback. Inspired by the recent advancements made by Generative Pre-trained Transformer (GPT) models, we conducted a study to examine the extent to which GPT models hold the potential to advance the existing knowledge of LA-supported feedback systems towards improving the efficiency of feedback provision. Therefore, our study explored the ability of two versions of GPT models – i.e., GPT-3.5 (ChatGPT) and GPT-4 – to generate assessment feedback on students' writing assessment tasks, common in higher education, with open-ended topics for a data science-related course. We compared the feedback generated by GPT models (namely GPT-3.5 and GPT-4) with the feedback provided by human instructors in terms of readability, effectiveness (content containing effective feedback components), and reliability (correct assessment on student performance). Results showed that (1) both GPT-3.5 and GPT-4 were able to generate more readable feedback than human instructors, (2) GPT-4 outperformed GPT-3.5 and human instructors in providing feedback containing information about effective feedback dimensions, including feeding-up, feeding-forward, process level, and self-regulation level, and (3) GPT-4 demonstrated higher reliability of feedback compared to GPT-3.5.

Keywords: Assessment Feedback, Learning Analytics, Feedback Generation, Generative Pre-Trained Transformer, Feedback Effectiveness, Automated Feedback System

Email addresses: wei.dai1@monash.edu (Wei Dai),
yi-shan.tsai@monash.edu (Yi-Shan Tsai), jiongh.lin@gmail.com
(Jionghao Lin), dragan.gasevic@monash.edu (Dragan Gašević),
guanliang.chen@monash.edu (Guanliang Chen)

Assessing the Proficiency of Large Language Models in Automatic Feedback Generation: An Evaluation Study

Abstract

Assessment feedback is important to student learning. Learning analytics (LA) powered by artificial intelligence exhibits profound potential in helping instructors with the laborious provision of feedback. Inspired by the recent advancements made by Generative Pre-trained Transformer (GPT) models, we conducted a study to examine the extent to which GPT models hold the potential to advance the existing knowledge of LA-supported feedback systems towards improving the efficiency of feedback provision. Therefore, our study explored the ability of two versions of GPT models – i.e., GPT-3.5 (ChatGPT) and GPT-4 – to generate assessment feedback on students' writing assessment tasks, common in higher education, with open-ended topics for a data science-related course. We compared the feedback generated by GPT models (namely GPT-3.5 and GPT-4) with the feedback provided by human instructors in terms of readability, effectiveness (content containing effective feedback components), and reliability (correct assessment on student performance). Results showed that (1) both GPT-3.5 and GPT-4 were able to generate more readable feedback than human instructors, (2) GPT-4 outperformed GPT-3.5 and human instructors in providing feedback containing information about effective feedback dimensions, including feeding-up, feeding-forward, process level, and self-regulation level, and (3) GPT-4 demonstrated higher reliability of feedback compared to GPT-3.5.

Keywords: Assessment Feedback, Learning Analytics, Feedback Generation, Generative Pre-Trained Transformer, Feedback Effectiveness, Automated Feedback System

1. Introduction

Assessment feedback is important to student learning. Quality feedback justifies grades, clarifies students' strengths and areas needing improvement, and develops students' learning skills. The limited quality of feedback may demotivate students' understanding and uptake of feedback, harm student-instructor relationships, and thus impact student success (Middleton et al., 2023). However, due to the expansion of higher education, instructors often express that they generally do not have the capacity to offer constructive comments for a large cohort of students (Henderson et al., 2019). Many students in higher education institutions have complained about the feedback they received concerning their performance, and expressed a desire for constructive assessment feedback (Mulliner and Tucker, 2017; Henderson et al., 2019).

Learning analytics (LA) is an interdisciplinary field where diverse disciplines such as education, psychology, computer science and statistics intertwine with each other to generate knowledge for the improvement of learning and teaching (Chatti et al., 2012). The emergence of the field of LA exhibits profound potential in helping instructors with the provision of personalised and real-time feedback at scale by setting up rules according to students' behaviour data on the learning management system (Arthars et al., 2019; Pardo et al., 2018). For instance, On-Task (Pardo et al., 2018) is one of the LA-supported feedback systems that enables instructors to deliver feedback at scale by setting up conditional rules about students' learning activities such as lesson attendance, risk of failing a course, and academic performance. Although these LA-supported feedback systems

alleviate instructors' workload in feedback provision to some extent, they still rely on instructors to assess students' performance and formulate personalised feedback, which poses a risk that students may receive feedback of sub-optimal quality due to instructors' limited capacity. This substantial reliance on instructors to oversee the feedback process is incongruent with the contemporary conceptualisation of feedback which positions students as having a pivotal role in managing their learning, and focuses on developing student feedback literacy, i.e., "the understandings, capacities and dispositions needed [by the learner] to make sense of information and use it to enhance work or learning strategies" (Carless and Boud, 2018, p.1316). Therefore, LA-supported systems for supporting the provision of feedback necessitate further exploration and investigation, especially in the growing power of generative artificial intelligence (Yan et al., 2024).

We posit that Generative Pre-trained Transformer (GPT) models hold the potential to advance the existing knowledge of LA-supported feedback systems towards improving the efficiency of feedback provision Yan et al. (2023). The adoption of self-attention mechanisms enables GPT models to handle long-range dependencies, and the fine-tuning on a large amount of conversational data allows it to generate coherent and contextually relevant answers. GPT models have shown their potential for feedback generation in the literature, such as generating corrective feedback on students' coding assignments (MacNeil et al., 2022; Pankiewicz and Baker, 2023), responding to students' posts on MOOC discussion forums (Li and Xing, 2021), and evaluating the use of effective praise, e.g., praising stu-

dents' effort, acknowledging students' learning process instead of learning outcomes, in the feedback provided by instructors (Hirunyasiri et al., 2023). For instance, GPT-2 was adopted to give feedback on students' posts published on MOOC discussion forums (Li and Xing, 2021) and achieved comparable performance on evaluation metrics of word perplexity, readability, and coherence to that of human experts. However, limited studies have applied GPT models to generate assessment feedback on writing tasks with open-ended topics such as students' project proposals. This kind of writing assignment on open-ended tasks is common in higher education yet instructors often struggle to deliver comprehensive feedback for each student in large enrolment classes (Beckman et al., 2021). To the best of our knowledge, only one study conducted by Dai et al. (2023) adopted GPT-3.5 to generate assessment feedback on writing tasks with open-ended topics. They compared the quality of feedback generated by GPT-3.5 and feedback written by instructors based on the theoretical framework of effective feedback developed by Hattie and Timperley (2007). Results showed that effectiveness of GPT-3.5 feedback is inferior to that of feedback written by instructors (Hirunyasiri et al., 2023).

Inspired by the recent advancements made by GPT-4 in various tasks such as solving medical challenge problems (Nori et al., 2023) and detecting praise in tutoring dialogues (Hirunyasiri et al., 2023), the current study extended the investigation into the capabilities of GPT models by comparing the quality of feedback generated by GPT-3.5 with its advanced model, i.e., GPT-4, on the same task as examined by Dai et al. (2023). We measured the quality of feedback generated by GPT models in terms of their readability, effectiveness, and reliability. Providing readable feedback is pivotal to facilitating students' understanding of the information conveyed, thereby promoting their actions for improvement (Evans, 2013). We evaluated the readability of feedback by focusing on examining the grammar, word choice and coherence. Effective feedback is paramount as it guarantees that the information conveyed has the greatest impact on student learning (Hattie and Timperley, 2007). We assessed the effectiveness of feedback by using a well-known theoretical framework for feedback proposed by Hattie and Timperley (2007) and analysing the presence of effective feedback components in feedback generated by GPT-3.5, GPT-4, and instructors. We define the *reliability* of machine-generated feedback as the agreement between machine-generated feedback and human-generated feedback in terms of feedback polarity (i.e., *Positive* if the feedback is provided to acknowledge students' performance or *Negative* if it suggests areas for improvement; Dai et al., 2023) by viewing human-generated feedback as the ground truth. Reliable feedback, which accurately reflects the quality of students' performance, is crucial in fostering a sense of trust between students and feedback providers, consequently enhancing student engagement in learning (Leighton and Bustos Gómez, 2018). Specifically, this study was guided by the following three research questions:

RQ1 To what extent is the feedback generated by GPT models readable in comparison to human-produced feedback?

RQ2 To what extent does the GPT-generated feedback contain effective feedback components to guide student learning in comparison to human-produced feedback?

RQ3 To what extent is the feedback generated by GPT models reliable when assessing students' performance in comparison to human-produced feedback?

2. Literature Review

2.1. Effective feedback design

Since feedback plays a crucial role in the learning process (Lin et al., 2023a,b; Dai et al., 2023), feedback researchers have been working on developing theoretical models that explain how feedback affects learning and what principles contribute to effective feedback design. For example, Nicol and Macfarlane-Dick (2006) identified seven principles of good feedback practice from the research literature on formative assessment, which support self-regulation from the cognitive, behavioural and motivational aspects. Hattie and Timperley (2007) proposed a feedback model that categorized effective feedback into three dimensions (i.e., *feeding up*, *feeding back* and *feeding forward*) and four-level focuses (i.e., *task*, *process*, *self-regulatory*, and *self*). Specifically, feeding up information tells the student where they are going, which is related to the judgment about the student's attainment of the learning goal (e.g., "*In this assignment, your goal is to clearly describe the business benefits of the project. You achieved the goal.*"). Feeding back information answers how the student is going and often consists of assessment comments on student performance on a specific part of the task (e.g., "*Your report has a very clear structure.*"). Feeding forward information addresses where the student is going next, which usually includes future improvements on the task. Feedback on *task* normally contains corrective information that indicates how well a task is performed (e.g., "*The interpretation of this machine learning model is incorrect.*"). Feedback on *process* is primarily aimed at suggesting strategies for completing the task (e.g., "*This page may make more sense if you use the strategies we talked about earlier.*"). Feedback on self-regulation addresses how students monitor their learning (e.g., "*You already know the key features of the opening of an argument. Check to see whether you have incorporated them in your first paragraph.*"), and feedback on self is about personal evaluations (e.g., "*You are a great student.*", "*Well done!*") (Hattie and Timperley, 2007). As principles proposed by Nicol and Macfarlane-Dick (2006) are considered too general and unsuitable for analysing textual feedback (Cavalcanti et al., 2020), and the feedback model proposed by Hattie and Timperley (2007) has been adopted widely in previous research to analyse textual feedback (Cavalcanti et al., 2020; Lin et al., 2023a; Osakwe et al., 2022), our study used the latter one to assess the effectiveness of feedback (RQ2).

2.2. Learning analytics-supported feedback

Feedback, as an essential component in guiding student learning, has gained much attention from LA researchers who

aim to enhance the quality and efficiency of feedback by using learning analytics technologies. OnTask (Pardo et al., 2018) is a feedback provision system where instructors can interpret the results about students' learning progress generated by automated data analytic methods and set appropriate actions to personalise feedback for students at scale. For example, if students' activity completion is over 70%, instructors can customise feedback for this group of students as the following, "It looks like you have been actively engaging in the activities in the course. Well done!" (Lim et al., 2021). Azcona et al. (2019) developed a predictive model to predict students' performance on a programming course based on their behavioural logs on courseware and demographics. For groups of students with different levels of predicted performance, personalised feedback emails regarding their engagement with courseware were designed by human experts and sent to students. Quantitative analysis showed that students gained learning improvement after receiving the feedback emails and survey results indicated that students' motivation and engagement increased because of the LA-supported feedback emails (Azcona et al., 2019). The Student Relationship Engagement System (SRES) (Arthars et al., 2019) allows instructors to import electronic data about student learning engagement and performance data, and conduct analysis to obtain students' learning patterns, based on which instructors can provide personalised support to students such as sending emails about student learning progress to students. These LA-supported feedback systems typically provide feedback based on pre-defined rules about student learning progress, which encounter limitations when applied to the automation of assessment feedback for writing tasks with open-ended topics such as essay assignments and project proposals. These tasks are gaining prominence in higher educational settings due to their role in developing students' self-regulated learning skills (Beckman et al., 2021). However, pre-defining sample answers or common errors for these types of tasks is challenging as student answers can be very diverse (Beckman et al., 2021). Moreover, the existing LA-supported feedback systems substantially rely on instructors to create the personalised feedback, which does not guarantee the quality of the feedback disseminated to students. Hence, it is imperative to pursue innovative solutions for advancing LA-supported feedback systems towards the goal of delivering quality feedback, especially for open-ended tasks.

2.3. Large language models for educational feedback

We posit that GPT-based large language models (LLMs) have the potential to augment the efficiency of feedback provision, due to the encouraging outcomes demonstrated by GPT-based LLMs in tasks associated with feedback generation. MacNeil et al. (2022) explored the ability of GPT-3 model to give feedback on the correctness of beginner programmers' code snippets. Results revealed that GPT-3 was able to provide feedback on the time complexity of the code, identify common mistakes, and summarise the goal of the code (MacNeil et al., 2022). A similar application of GPT-based LLMs is proposed by Pankiewicz and Baker (2023) who utilised the GPT-3.5 model to generate feedback on students' programming assignments.

Comparative experiments were conducted to assess the impact of feedback generated by the GPT-3.5 model on students' code submissions. The results indicated that students who received feedback generated by the GPT-3.5 model achieved a higher percentage of successful submissions than those who did not receive GPT-3.5 generated feedback (Pankiewicz and Baker, 2023). Instead of providing feedback to students, Hirunyasiri et al. (2023) aimed at offering timely feedback to human tutors specifically on their performance on giving effective praise to students. They adopted the latest version of GPT models, i.e., GPT-4, and experimental results on synthetic tutoring dialogues showed that GPT-4 performed better than human annotators in detecting specific and immediate praise. Despite that these studies have presented the feasibility of GPT-based models on feedback generation, none of them has attempted to generate assessment feedback on writing tasks with open-ended topics such as students' project proposals. This kind of writing assignment on open-ended tasks is common in higher education yet instructors often struggle to deliver comprehensive feedback for each student in large enrolment classes (Beckman et al., 2021).

A recent study (Dai et al., 2023) explored the capability of GPT-3.5 model (ChatGPT) to generate assessment feedback for students' project writing proposals in higher education and found that GPT-3.5 could provide more readable feedback than human assessors. Nonetheless, when evaluated against the presence of effective components as suggested in the feedback model by Hattie and Timperley (2007), the effectiveness of GPT-3.5 generated feedback was inferior to that of feedback composed by instructors, despite a considerable number of effective feedback components (such as task-level and process-level feedback) observed in GPT-3.5 generated feedback. Moreover, GPT-3.5 did not achieve high agreement with human instructors in terms of feedback polarity (i.e., *Positive* if the feedback is provided to acknowledge students' performance or *Negative* if it suggests areas for improvement). Unreliable assessment feedback provided by the machine, i.e., positive feedback on poor performance or negative feedback on good performance, may inadvertently convey misleading information to students about their learning and thus discourage student achievement (Dawson et al., 2019; Dai et al., 2023).

GPT-4 exhibited superior performance compared to GPT-3.5 in various educational tasks including solving medical challenge problems (Nori et al., 2023) and detecting praise in tutoring dialogues (Hirunyasiri et al., 2023). This enhanced performance is attributed to a significant increase in the number of model parameters and the volume of training data (Nori et al., 2023). Therefore, the current study aimed to incorporate this advanced model to further explore the capabilities of the GPT models on the task of feedback generation.

3. Methods

3.1. Data Collection

3.1.1. Instructor-generated feedback collection

Our study obtained ethics approval from ANONYMOUS University under project number [BLINDED]. We retrieved

Table 1: Statistics for report length counted by the number of words.

Min	Median	Max	Mean	Std
203	800	1811	786.87	241.94

the dataset from a postgraduate-level course teaching introductory data science skills that was delivered during the second semester of 2022. In this course, students were required to propose a data science project relevant to a specific domain including medical, financial, and educational systems, and submit a project writing proposal for marking. The proposals should include two sections (project description and business model) about the application of data science techniques in a specific domain. Instructors evaluated the submitted proposal and provided textual feedback for each student according to the following five aspects specified in the marking rubric: i) clear description of the goals of the project (**Goal**), ii) appropriateness of the topic to data science (**Topic**), iii) clear description of the business benefits (**Benefit**), iv) novelty/creativity (**Novelty**), and v) overall clarity of the report (**Clarity**).

After removing the student records without feedback, we finally obtained 103 students’ proposal reports and the associated instructor-generated feedback. Note that we removed the identifiable personal information of students both in reports and feedback for the protection of privacy. The length of students’ proposal reports varied from 203 to 1,811 words, which falls within GPT’s maximum input capacity. Descriptive statistics of the length of proposal reports, and the length of instructor-generated feedback are presented in Table 1 and Table 2, respectively.

3.1.2. Feedback Generation by GPT-3.5 and GPT-4

We accessed GPT-3.5 and GPT-4 via the interface of ChatGPT developed by OpenAI Aydın and Karaarslan (2022). GPT models are able to generate responses from seeing a prompt describing the task (i.e., an instruction or a query written in natural language by a user for the model to execute). In the current study, the task for GPT models was to generate textual feedback on students’ proposal reports in terms of five assessment aspects. Thus, we designed the prompt for GPT-3.5 and GPT-4 as follows, “Please give feedback on the following text in terms of a clear description of the goals of the project, appropriateness of the topic to data science, a clear description of the business benefits, novelty/creativity and overall clarity of the report. <INSERT THE TEXT OF A REPORT>”. For each student’s proposal, we inserted the text of their proposal report into the prompt and submitted it to GPT-3.5 and GPT-4 to obtain generated feedback. To prevent student assignments from being collected by models, we unchecked the option of “Chat history & training” in the settings of the model interface. The statistics of the length of feedback generated by GPT-3.5 and GPT-4 are shown in Table 2.

Table 2: Statistics for feedback length counted by the number of words.

Feedback	Min	Median	Max	Mean	Std
Instructor	6	51	143	57.34	30.60
GPT-3.5	102	160	270	166.44	34.81
GPT-4	275	361	522	369.88	51.81

3.2. Measuring readability

To answer RQ1, we adopted a widely-used measure – i.e., readability (van der Lee et al., 2021; Jia et al., 2022) – to examine the quality of the machine-generated text in grammar, word choice, and coherence. In line with the process of evaluating readability in written products (Jia et al., 2022), we invited three human raters, and each of them was asked to score all pieces of feedback generated by instructors, GPT-3.5 and GPT-4 using a five-point scale where: (i) 0 denoted *Incomprehensible*, (ii) 1 *Not fluent and incoherent*, (iii) 2 *Somewhat fluent but incoherent*, (iv) 3 *Fluent but somewhat incoherent*, and (v) 4 *Fluent and coherent*. As the evaluation for text readability varied from one individual human rater to another, we calculated the average score of the three human raters as the final metric of readability for each piece of feedback.

3.3. Measuring feedback effectiveness

To answer RQ2, we used a well-known theoretical framework for feedback proposed by Hattie and Timperley (2007) and analysed the presence of effective feedback components in the feedback generated by GPT-3.5, GPT-4, and instructors. Please refer to Section 2.1 for the details of this framework. We recruited two experts to annotate feedback by using the three dimensions and four levels proposed by Hattie and Timperley (2007). After a pre-training session about annotation rules, each expert annotated 309 pieces of feedback in our dataset (103 pieces of instructor-generated feedback, 103 feedback generated by GPT-3.5, and 103 feedback by GPT-4). To avoid introducing bias towards human instructors or GPT models, we did not inform the experts which feedback was written by instructors and which was generated by GPT-based models. Two experts achieved high agreement scores between their annotations, as shown in Table 3. For feedback received inconsistent labels from two experts, a third expert was invited to make final decision. We used the labels with inconsistencies resolved by the third expert for analysis.

3.4. Measuring the reliability of feedback generated by GPT-based models

Providing students with reliable feedback regarding their assignment performance is of great value for on-going learning experience (LeClercq, 1999). When students accomplish the set goal, the instructor may acknowledge the students’ effort in the feedback. When students perform undesirably, the instructor may indicate the areas that they should further improve. In this paper, we use *Polarity* to denote whether the feedback is given to acknowledge students’ effort (*Positive*) or indicate the

areas that they should further improve (*Negative*). If a feedback generator was unable to give reliable feedback that accurately indicated how well students’ performance was – i.e., positive feedback on poor performance or negative feedback on good performance – the generated feedback may have inadvertently misled the student and negatively impacted learning. Hence, we investigated the reliability of GPT-generated feedback by measuring the agreement between GPT models and human instructors in terms of feedback polarity.

Two human annotators were recruited to identify the feedback polarity. Each piece of feedback from either instructor or GPT-based models was generated by assessing the students’ reports based on five aspects (i.e., Goal, Topic, Benefit, Novelty and Clarity). Thus, we identified the polarity of feedback for each of these aspects. Specifically, if feedback was given to acknowledge a student’s effort on a specific aspect, then the annotator marked it with *Positive*, whereas if feedback was given to indicate that the student needs to improve on a specific aspect, the annotator marked it with *Negative*. If the feedback did not contain any comments on a specific aspect, then the annotator marked it with *None*. For each piece of instructor-generated or GPT-generated feedback, we obtained five labels, and each of them indicates the feedback polarity (i.e., “Positive”, “Negative” or “None”) on each of the five assessment aspects.

We then evaluated the ability of GPT models on generating feedback for each assessment aspect with accurate polarity. By doing so, we calculated precision, recall and F_1 score (shown in Equations 1, 2 and 3), three commonly-used metrics for reliability of multi-class classification tasks, by regarding three feedback polarities as three classes, the labels of instructor feedback as the ground truth, and labels of feedback generated by GPT-based models as predicted classes, as the purpose of this study was to evaluate the feasibility of using GPT-based models to support human educators in feedback provision. For example, supposing that GPT-3.5 gave positive feedback on the aspect of **Topic** to four reports, while only two of them obtained positive feedback from the instructor on the same aspect, then the precision was $2/4$ (i.e., Precision = 0.50). Supposing that the instructor gave positive feedback on the aspect of **Topic** to 6 reports in total, the recall was calculated as $2/6$ (0.33). F_1 score was calculated as 0.4, i.e., the harmonic mean of the precision and recall scores, as shown below.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (1)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (2)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

When we annotated feedback polarity, we found that human instructors did not comment on some of the assessment

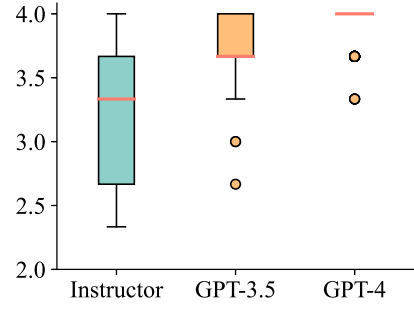


Figure 1: The distribution of feedback readability.

aspects defined in the marking rubric (i.e., Goal, Topic, Benefit, Novelty and Clarity), while GPT models offered comprehensive feedback (i.e., feedback on more assessment aspects in the marking rubric). For example, for the assignment submitted by a specific student, the human instructor provided comments only on the assessment aspects of **Goal** and **Topic**. However, GPT-generated feedback on the same assignment contained comments on three assessment aspects including **Goal**, **Topic** and **Benefit**. In this case, for the assessment aspect of **Benefit**, we did not have the ground truth, i.e., the human instructor’s feedback on this aspect, to compare the identified feedback polarity on **Benefit** between a GPT model and the human instructor. Therefore, we need to further evaluate the reliability of this group of GPT-generated feedback comments. We first extracted all the GPT-generated feedback comments on the assessment aspects that human instructors did not comment on from our dataset. Then, we asked a human assessor to assess the assignment submitted by the student who received this feedback and then answer the question: “Do you agree with this feedback regarding the student’s assignment?”. The human assessor selected the answer from the following three options: i) Agree; ii) Partially agree, and iii) Disagree. As a piece of feedback could contain multiple sentences of comments on different assessment aspects and the human assessor agrees with some of these sentences, we provided the option of “Partially agree” for the human assessor as an option to choose.

4. Results

4.1. RQ1

Table 4 presents the readability scores between GPT-3.5, GPT-4, and instructor feedback. Statistical difference of readability between human, GPT-3.5, and GPT-4 are shown in Table 5. We can see that feedback generated by both GPT-based models was significantly more readable than instructor-generated feedback. Figure 1 shows the distribution of readability scores for the feedback provided by instructors and two GPT models. The distribution indicates that the majority of readability scores on GPT-generated feedback was rated between 3.75 to 4.0, whereas most instructor feedback (over 75%) was below the readability score of 3.75 with a higher standard deviation than those of the GPT-based models.

Table 3: Agreement scores and Cohen’s κ between two annotators.

Categories	Agreement score			Cohen’s κ		
	Instructor	GPT-3.5	GPT-4	Instructor	GPT-3.5	GPT-4
Five assessment aspects						
Goal	1.00	0.98	0.99	1.00	0.87	0.98
Topic	0.98	0.98	1.00	0.93	0.85	1.00
Benefit	0.99	0.95	0.98	0.98	0.83	0.96
Novelty	0.98	0.88	0.98	0.98	0.81	0.96
Clarity	0.97	0.92	0.96	0.94	0.82	0.89
Seven effective feedback components						
Feeding up	1.00	1.00	0.98	1.00	None ²	0.79
Feeding back	0.97	1.00	1.00	0.99	1.00	All ¹
Feeding forward	0.99	0.98	0.99	0.95	0.96	0.80
Task	1.00	1.00	1.00	All ¹	All ¹	All ¹
Process	0.95	0.92	0.98	0.85	0.84	0.74
Self-regulation	0.96	1.00	0.94	0.81	None ²	0.77
Self	0.98	1.00	1.00	0.95	None ²	None ²

¹ All feedback contained comments in this category.² No feedback contained comments in this category.

Table 4: Statistics for feedback readability.

Feedback	Min	Median	Max	Mean	Std
Instructor	2.33	3.33	4	3.21	0.52
GPT-3.5	2.67	3.67	4	3.76	0.28
GPT-4	3.33	4	4	3.91	0.18

Within GPT-based models, GPT-4 performed significantly better than GPT-3.5 did in generating readable feedback. By scrutinizing the content of feedback generated by instructors, GPT-3.5 and GPT-4 (Feedback examples on the same student assignment extracted from our dataset can be seen in Table 6), we found some differences between these three types of feedback. Feedback from the instructor consisted mostly of simple expressions about the evaluation of student performance, while GPT-generated feedback usually contained a brief summary of students’ works and the evaluation feedback followed by an explanation of how this feedback was given. Comparing GPT-3.5 and GPT-4, we can see most of the feedback by GPT-4 was formatted using bullet points to clarify the comments for each aspect in the marking rubric.

4.2. RQ2

Table 7 reveals that human assessors provided **feeding-up** information (i.e., information related to the judgment about the student’s attainment of the learning goal) only in six pieces of feedback in the used dataset. GPT-3.5 did not offer any of this information in its generated feedback. In contrast, GPT-4 generated **feeding-up** information in almost every piece of feedback (94.17%). This difference was not observed in the provision of **feeding-back** information (i.e., information related to the assessment on student performance on a specific part of the task), as nearly 100% feedback generated by human assessors,

GPT-3.5, and GPT-4 contained this information. **Feeding-forward** information (i.e., information related to suggestions for future improvements on the task) was detected in 93 pieces of feedback written by human assessors. The same information was detected less frequently in the feedback generated by GPT-3.5. It is worth noting that GPT-4 performed the best in generating **feeding-forward** information (existing in 100 out of 103 pieces of the generated feedback).

We observed **task** level information in every piece of feedback generated by human assessors, GPT-3.5 and GPT-4. As for **process** level information, we detected it in 80% of feedback from human assessors, compared to 57% of feedback generated by GPT-3.5. Notably, this information was detected 100% in the feedback generated by GPT-4. Information on **self-regulation** was detected in only 11 (out of 103) pieces of feedback written by human assessors. This information was not observed in GPT-3.5-generated feedback, but in 18 pieces of feedback generated by GPT-4, slightly higher than human-generated feedback. Finally, the **self** level information was detected in 25 pieces of human feedback, but not in GPT-3.5 or GPT-4 generated feedback at all.

Overall, GPT-4 appeared to have outperformed both GPT-3.5 and human instructors in terms of producing balanced feedback that consistently covers most of the seven effective feedback components except for the self-regulation and self levels. It is worth noting that despite the relatively lower presence of information on self-regulation, GPT-4 still outperformed GPT-3.5 and human instructors in this aspect. However, it is clear that both GPT models placed less emphasis on self-level feedback, which can be an important feedback element to motivate students, especially through affirming their performance (Hattie and Timperley, 2007).

Table 5: Statistical difference between the three groups using t-test and Cohen's d. Significant values are denoted with "***"

Groups	t-test ¹			Cohen's d		
	t-score	p-value	adjusted p-value ²	Mean of X	Mean of Y	effect size
Instructor vs GPT-3.5	-9.50	1.07e-15	3.20e-15***	3.21	3.76	-1.31
Instructor vs GPT-4	-12.56	2.00e-22	5.99e-22***	3.21	3.91	-1.79
GPT-3.5 vs GPT-4	-4.79	5.75e-06	1.73e-05***	3.76	3.91	-0.64

¹t-tests were conducted as post hoc tests after ANOVA test (f-value = 109.76 and p-value = 1.17e-36).

²p-value was adjusted using Bonferroni correction.

Table 6: Feedback examples on the same assignment generated by human instructor, GPT-3.5 and GPT-4. Sentences that summarise the assignment are omitted.

Feedback generator	Feedback example
Instructor	<i>"The description is clear. Good choice of topic. Good description of the business model. A novel creative idea. Clear and well structured report."</i>
GPT-3.5	<i>"The text provided describes a project that aims to...The text also mentions...Overall, the goals of the project are clear and the use of data science in analyzing job listings is appropriate. The potential business benefits of the project, such as reducing costs and identifying the requirements for different data science roles, are also clearly described. The novelty and creativity of the project may depend on the specific methods and analyses used in the project. The overall clarity of the report could be improved by providing more detail on the methods and analyses that will be used, as well as the expected results and implications of the project."</i>
GPT-4	<p><i>"Overall, the report has a promising project idea but needs substantial editing for clarity, precision, and professionalism. Here's a breakdown of the feedback based on your requested criteria:</i></p> <ul style="list-style-type: none"> <i>• Clear Description of the Goals of the Project: The goal of the project is generally clear. It aims to...However, these goals could be presented more concisely and clearly.</i> <i>• Appropriateness of the Topic to Data Science: The topic is very relevant to data science as it involves the analysis of job listings, a type of structured data. The proposal mentions various analytical tasks, such as...</i> <i>• Clear Description of the Business Benefits: The report mentions the benefits for students, companies, and the government, but the explanations are somewhat vague and confusing. The report suggests that...It's not clear what costs and problems it refers to, and how exactly they would be reduced. The report also suggests..., but this concept is not well defined or explained.</i> <i>• Novelty/Creativity: The project is reasonably novel in focusing specifically on...However, the creativity of the proposed methods is hard to assess based on the report. The report does not elaborate on the specific techniques or algorithms to be used in the analysis.</i> <i>• Overall Clarity of the Report: The overall clarity of the report could be improved. The report contains several long, run-on sentences, which make it hard to follow. The writing style could be more academic and professional. There are also some spelling and grammar errors. The structure of the report could be clearer, with distinct sections for the project goals, the data to be used, the analysis methods, the expected benefits, and the potential challenges. The references section could also be formatted more professionally."</i>

Table 7: The comparison of the distribution of seven effective feedback components between the feedback provided by the human instructor, generated by GPT-3.5 and GPT-4. For each component, ANOVA tests were performed to measure the statistical difference between the three feedback generators. Subsequent post hoc comparisons were conducted using t-tests. Pairs of feedback generators that demonstrated significant differences (p-value ≤ 0.05) are listed in the final column. For example, “GPT-4 vs {Instructor,GPT-3.5}” signifies that significant differences were identified between GPT-4 and Instructor, as well as between GPT-4 and GPT-3.5. The term “All” indicates that each pair of feedback generators exhibited a significant difference, whereas “None” denotes the absence of significant difference among all three feedback generators.

Components	Instructor		GPT-3.5		GPT-4		Significant results
	Quantity	Frequency	Quantity	Frequency	Quantity	Frequency	
Feeding up	6	5.83%	0	0%	97	94.17%	GPT-4 vs {Instructor, GPT-3.5}
Feeding back	101	98.06%	102	99.03%	103	100%	None
Feeding forward	93	90.29%	63	61.17%	100	97.09%	GPT-3.5 vs {Instructor, GPT-4}
Task	103	100%	103	100%	103	100%	None
Process	82	80%	57	55%	100	97.09%	All
Self-regulation	11	11%	0	0%	18	17.48%	GPT-3.5 vs {Instructor, GPT-4}
Self	25	24%	0	0%	0	0%	Instructor vs {GPT-3.5, GPT-4}

Table 8: Polarity of feedback generated by human instructors, GPT-3.5, and GPT-4 according to the five assessment aspects in the marking rubric. The precision, recall and F1 for measuring the agreement between GPT models and human assessors in terms of feedback polarity are included.

Aspects	Polarity	Quantity			Precision		Recall		F1	
		Instructor	GPT-3.5	GPT-4	GPT-3.5	GPT-4	GPT-3.5	GPT-4	GPT-3.5	GPT-4
Goal	Positive	5	95	74	0.04	0.05	0.80	0.80	0.08	0.10
	Negative	40	3	29	0.33	0.62	0.03	0.45	0.06	0.52
	None	58	5	0	0.40	0.00	0.03	0.00	0.06	0.00
Topic	Positive	86	95	99	0.84	0.85	0.93	0.98	0.88	0.91
	Negative	5	0	3	0.00	0.67	0.00	0.40	0.00	0.50
	None	12	8	1	0.13	0.00	0.08	0.00	0.10	0.00
Benefit	Positive	19	85	64	0.20	0.22	0.90	0.74	0.33	0.34
	Negative	59	5	37	0.60	0.73	0.05	0.46	0.09	0.56
	None	25	13	2	0.31	0.00	0.16	0.00	0.21	0.00
Novelty	Positive	31	52	64	0.35	0.44	0.58	0.90	0.44	0.59
	Negative	19	22	39	0.27	0.33	0.32	0.68	0.29	0.44
	None	53	29	0	0.59	0.00	0.32	0.00	0.41	0.00
Clarity	Positive	24	77	76	0.21	0.28	0.67	0.88	0.32	0.42
	Negative	11	17	27	0.29	0.19	0.46	0.46	0.36	0.27
	None	68	9	0	0.44	0.00	0.06	0.00	0.11	0.00

Table 9: “No. of feedback” denotes the number of feedback generated by GPT-3.5 or GPT-4 on the assessment aspect that was not commented on by the original human assessors. “Quantity” denotes the number of feedback that the human instructor agreed (partially agreed/disagreed) with. “Percent” was calculated as “Quantity” divided by “No. of feedback”.

Model	No. of feedback	Agree		PartAgree		Disagree	
		Quantity	Percent	Quantity	Percent	Quantity	Percent
GPT-3.5	60	44	73.33%	9	15%	7	11.67%
GPT-4	101	91	90.10%	7	6.93%	3	2.97%

4.3. RQ3

The results for answering RQ3 are presented in Table 8. In terms of generating feedback according to the marking rubric, GPT-4 provided comments on almost all the five assessment aspects in every single feedback except for a relatively low number of feedback that failed to cover the **Topic** ($n = 1$) and **Benefit** ($n = 2$) aspects. Compared to GPT-4, GPT-3.5 performed slightly worse in covering these five aspects. In particular, 29 pieces of GPT-3.5 feedback missed comments on the **Novelty** aspect. Compared to GPT-4 and GPT-3.5, human instructors appeared to have missed comments on these five aspects more frequently. In particular, over half of the students did not receive comments on **Goal**, **Novelty**, and **Clarity**.

By observing the quantity of feedback of different polarity, we found that both the instructor and GPT models gave a majority of students positive feedback on the aspect of **Topic**, i.e., 86 out of 103 positive feedback by the instructor, 95 out of 103 by GPT-3.5, and 74 out of 103 by GPT-4. On the other four aspects, students mostly received either negative feedback or empty feedback (i.e., feedback that did not contain any comments on a specific aspect, which was denoted as None in Table 8) from the instructor, in comparison with GPT models which generated more positive feedback than the instructor.

By observing values of precision, recall and F_1 (introduced in Sec. 3.4) in Table 8, we can see that among the five assessment aspects, the GPT models achieved the highest agreement score with human instructors (precision: 0.85, recall: 0.98, F_1 : 0.91) in positive polarity for the aspect of **Topic** (note that we considered human-generated feedback as the ground truth in terms of the detected feedback polarity). This suggests that both the instructor and GPT models gave a majority of students positive feedback on the aspect of **Topic**, i.e., 86, 95, and 99 out of the 103 feedback by human instructors, GPT-3.5, and GPT-4, respectively, contained positive comments on this aspect. Although GPT models did not achieve high agreement with instructors on the other four assessment aspects, we can observe an increase of F_1 for GPT-4 compared to GPT-3.5 except for negative polarity for the aspect of **Clarity**.

According to Table 9, we can see that 60 out of 103 pieces of feedback generated by GPT-3.5 had comments related to the assessment aspects that were not commented on by the original human instructors, and 73.33% of these pieces of feedback obtained agreement from the human assessor. In comparison to GPT-3.5, more feedback pieces generated by GPT-4 (101 out of 103 pieces of feedback) contained comments that were not commented on by the original human instructors, and 90.10% of these pieces of feedback obtained agreement from the human assessor. The proportion of the feedback that the human assessor disagreed with for both generators – i.e., GPT-3.5 and GPT-4 – was lower than those the human assessor agreed and partially agreed with. We extracted examples of feedback comments generated by GPT-3.5 and GPT-4 on the assessment aspects that were not commented on by the human instructor in Table 10

5. Discussion

5.1. Interpretation of the Results

The results of RQ1 showed that the GPT models were able to generate more readable feedback with greater consistency than human instructors. This is not surprising because the ability of LLMs to provide readable responses in various contexts has been demonstrated in the literature Li and Xing (2021); Patel et al. (2023). For example, Li and Xing (2021) indicated that GPT-2 could provide readable and coherent replies to human learners' posts on MOOC discussion forums to a similar extent compared to human instructors. Human instructors often struggle to consistently deliver quality feedback that meets students' expectations Weaver (2006). As investigated by Dai et al. (2022), there can be inconsistent delivery of polite feedback across students with different levels of performance and demographics. Existing LA-supported feedback systems mostly rely on human instructors to create feedback Pardo et al. (2018); Arthars et al. (2019), which risks delivering feedback of inconsistent quality. This could be mitigated by incorporating large language models such as GPT models in LA-supported feedback systems to help human instructors provide feedback of consistent readability to make students understand the feedback better and thus act upon the given feedback Evans (2013).

The results of RQ2 indicated that GPT-4 was able to generate feedback containing effective components more consistently than human instructors, based on the prominent theoretical framework proposed by Hattie and Timperley (2007). This is surprising as the evidence has previously been reported (Dai et al., 2023) that large language models such as the GPT-3.5 model performed less well than human instructors in providing feedback that covers effective components by using the same theoretical framework (Hattie and Timperley, 2007). Specifically, GPT-4 was superior to GPT-3.5 and even human instructors in providing feedback with more effective components, especially the feeding-up information, process-level and self-regulation level information. The high presence of feeding-up information in GPT-4 generated feedback boils down to the fact that GPT-4 explicitly included the assessment aspects in the marking rubric, i.e., i) clear description of the goals of the project, ii) appropriateness of the topic to data science, iii) clear description of the business benefits, iv) novelty/creativity, and v) overall clarity of the report, and connected them with student performance in the generated feedback to inform students of the attainment of learning goals, as marked in bold in the feedback examples provided in Table 6. By contrast, GPT-3.5 and human instructors very rarely referred to the assessment aspects in the provided feedback. Our study also found that human instructors provided less feedback at the **self-regulation** level compared to the other three levels (**task**, **process**, and **self**). This finding aligns with the findings previous reported in the literature (Osakwe et al., 2022; Cavalcanti et al., 2020).

The results of RQ3 in Table 8 indicated low precision, recall and F_1 for the GPT models in all assessment aspects except for the aspect of **Topic** in terms of the agreement on feedback polarity by considering the polarity of feedback generated by human instructors as the ground truth. This means that GPT

Table 10: Examples of feedback comments generated by GPT-3.5 and GPT-4 on the assessment aspects the human instructor did not comment on.

GPT-3.5

Example 1: “However, the text could be made more concise and clearer in some parts, such as the description of the project roles and the classification of content composers.”

Example 2: “In terms of novelty and creativity, the text could benefit from discussing any unique or innovative approaches to data analysis or prediction being used on the platform.”

Example 3: “However, it may be helpful to include more information on the specific data sources and methods that will be used in the analysis, as well as the expected outcomes and potential impact of the project. It would also be helpful to provide more detail on the real-world problem the project aims to solve and how the proposed solution will address it.”

GPT-4

Example 1: “However, a few improvements could be made for better clarity: The Abstract could be more precise about the specific goals and objectives of the study. The ‘Business Model’ section could be more directly related to Liverpool FC, and how data science contributes to each revenue stream. The report could be more explicit about the project’s specific objectives, methods, and expected outcomes, which is typical in a project proposal.”

Example 2: “The topic is appropriate for data science as it involves data collection, cleaning, and analysis, as well as the application of algorithms to make recommendations based on user needs and market demands. However, it would be beneficial to provide more detail about the types of data science techniques and methodologies to be used, for example, which machine learning algorithms might be relevant for this project.”

Example 3: “One suggestion is to elaborate more on the data science methods to be employed, particularly on the machine learning aspect. Which specific models or algorithms do you plan to use? Also, consider discussing how you would address potential privacy concerns related to collecting personal data like medical history.”

models could generate relatively reliable results in assessing whether the students performed well or poorly in generating a report that aligned with the topics defined for the assignment. However, GPT models could not offer a reliable assessment of student performance on the other aspects of the rubric. Table 8 shows that human instructors did not offer comprehensive feedback (i.e., feedback on each of the assessment aspects in the marking rubric) to students. This may be attributed to limited time human instructors have to provide feedback on students’ assignments. Therefore, they need to prioritise particular information they want to emphasise in their feedback. However, this feedback can be constrained, as human instructors might only comment on the assessment aspects that they noticed in the assignment within the limited time available. This is not applicable to GPT models, as they were asked to generate feedback on all the assessment aspects. This difference might explain the low precision, recall and F_1 in terms of feedback polarity. The results in Table 9 revealed that the human assessor agreed with the majority of feedback that GPT models generated on the assessment aspects that human instructors did not comment on. This observation suggests that GPT models possess the potential to assist human instructors in delivering feedback that covers the comprehensive aspects of student performance.

5.2. Implications

To promote the effective use of feedback, instructors need to pay more attention to the presence of effective components in the crafted feedback (Hattie and Timperley, 2007). According to the results of our study, GPT-4 yielded better performance in

generating feedback that contains effective feedback elements. Therefore, instructors can leverage GPT models to improve the quality of their feedback by promoting GPT models to include more feeding up, process level and self-regulation level information. The instructors can also use LLMs to improve their feedback skills (Markel et al., 2023). However, manually polishing feedback via ChatGPT interface for a large number of students can be time-consuming. Prompt design might be a challenging task for instructors without a background in engineering large language models. An alternative way is to encourage students to seek feedback from GPT models. As shown in our study, GPT models can generate more comprehensive feedback than human instructors and can detect issues in their work, which could be omitted by human instructors. Moreover, by positioning the student as the central actor in the feedback process, the skills required to evaluate their own work, seek feedback from multiple sources, and use feedback to improve, commonly known as *feedback literacy*, can be developed (Ryan et al., 2023; Carless and Boud, 2018). Given that LA can offer opportunities to trace students’ engagement with feedback (Lang et al., 2017), the integration of GPT models into LA-supported feedback systems can close a full feedback loop (which is always at the core of LA Clow (2012)). This loop aims not only to develop and enhance student feedback literacy by understanding students’ sense-making and action-taking processes, but also encourage students to actively engage with their feedback Carless and Boud (2018).

Despite the benefits brought by generative artificial intelligence such as GPT models in education, issues associated with

the incorporation of GPT models in feedback provision need to be carefully addressed. While GPT models are advanced, they are not infallible and may generate inaccurate or unreliable feedback information which could inadvertently mislead student learning and lead to negative emotions towards learning (Middleton et al., 2023). As revealed in our study, GPT models could not offer a reliable assessment of student performance on the aspects of the rubric except for **Topic**. Therefore, students can not solely rely on GPT-based tools to obtain feedback. Feedback from instructors, peers and other educational resources is necessary to ensure an accurate and reliable learning environment (Shute, 2008).

In the meantime, system developers and LA researchers who are seeking ways to incorporate LLMs into LA-supported feedback systems need to devise a way to validate and update GPT models regularly to improve the reliability of the generated feedback. A potential approach could be creating an iterative feedback loop where instructors and students can provide immediate responses to the feedback given by GPT models regarding its relevance to student assignment, and its reliability as a source of assessment information. This mechanism could help with immediate refinements and adaptations on GPT models, thereby contributing to the progressive enhancement of the models for the provision of reliable assessment feedback.

5.3. Limitations

Firstly, GPT models generated feedback in an unsupervised way, which could potentially influence the reliability of the generated assessment feedback in our study. Future efforts need to consider designing feedback samples that clearly and accurately describe student performance, and submit them to train GPT models to generate feedback with higher reliability. *Secondly*, we measured the effectiveness of feedback based on one single theoretical feedback framework proposed by Hattie and Timperley (2007), but the effectiveness of feedback goes beyond the content coverage of feedback (Yang and Carless, 2013; Carless and Boud, 2018; Middleton et al., 2023). Therefore, feedback models to facilitate relationships between instructors and students (Carless and Boud, 2018; Middleton et al., 2023), or to develop student feedback literacy (Yang and Carless, 2013) could be adopted to evaluate the quality of GPT-generated feedback. Future studies are also needed to compare the learning outcomes between students who receive GPT-generated feedback and those who receive instructor-written feedback to better evaluate the effectiveness of feedback based on LLMs. *Lastly*, the assignments and corresponding feedback were collected from the same course and for a single type of assessment task. It is important to examine the ability of GPT models for feedback generation for other courses in various subjects and other types of assessment tasks to determine the generalisation of the findings presented in this paper.

References

Arthars, N., Dollinger, M., Vigentini, L., Liu, D.Y.T., Kondo, E., King, D.M., 2019. Empowering teachers to personalize learning support: Case studies of

- teachers' experiences adopting a student-and teacher-centered learning analytics platform at three australian universities. Utilizing learning analytics to support study success , 223–248.
- Aydn, Ö., Karaarslan, E., 2022. OpenAI ChatGPT generated literature review: Digital twin in healthcare. Available at SSRN 4308687 .
- Azcona, D., Hsiao, I.H., Smeaton, A.F., 2019. Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. User Modeling and User-Adapted Interaction 29, 759–788.
- Beckman, K., Apps, T., Bennett, S., Dalgarno, B., Kennedy, G., Lockyer, L., 2021. Self-regulation in open-ended online assignment tasks: the importance of initial task interpretation and goal setting. Studies in Higher Education 46, 821–835.
- Carless, D., Boud, D., 2018. The development of student feedback literacy: enabling uptake of feedback. Assessment & Evaluation in Higher Education 43, 1315–1325.
- Cavalcanti, A.P., Diego, A., Mello, R.F., Mangaroska, K., Nascimento, A., Freitas, F., Gašević, D., 2020. How good is my feedback? a content analysis of written feedback, in: Proceedings of the 10th International Conference on Learning Analytics & Knowledge, pp. 428–437.
- Chatti, M.A., Dyckhoff, A.L., Schroeder, U., Thüs, H., 2012. A reference model for learning analytics. International Journal of Technology Enhanced Learning 4, 318–331.
- Clow, D., 2012. The learning analytics cycle: closing the loop effectively, in: Proceedings of the 2nd international conference on learning analytics and knowledge, pp. 134–138.
- Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.S., Gasevic, D., Chen, G., 2023. Can large language models provide feedback to students? a case study on ChatGPT, in: Proceedings of the 23rd IEEE International Conference on Advanced Learning Technologies, pp. 323–325.
- Dai, W., Tsai, Y.S., Fan, Y., Gašević, D., Chen, G., 2022. Measuring inconsistency in written feedback: A case study in politeness, in: Proceedings of the 23rd International Conference of Artificial Intelligence in Education, pp. 560–566.
- Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., Molloy, E., 2019. What makes for effective feedback: staff and student perspectives. Assessment & Evaluation in Higher Education 44, 25–36.
- Evans, C., 2013. Making sense of assessment feedback in higher education. Review of educational research 83, 70–120.
- Hattie, J., Timperley, H., 2007. The power of feedback. Review of Educational Research 77, 81–112.
- Henderson, M., Ryan, T., Phillips, M., 2019. The challenges of feedback in higher education. Assessment & Evaluation in Higher Education .
- Hirunyasiri, D., Thomas, D.R., Lin, J., Koedinger, K.R., Aleven, V., 2023. Comparative analysis of gpt-4 and human graders in evaluating praise given to students in synthetic dialogues. arXiv preprint arXiv:2307.02018 .
- Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., Gehringer, E., 2022. Insta-reviewer: A data-driven approach for generating instant feedback on students' project reports. Proceedings of the 15th International Conference on Educational Data Mining , 5–16.
- Lang, C., Siemens, G., Wise, A., Gasevic, D., 2017. Handbook of learning analytics. SOLAR, Society for Learning Analytics and Research New York.
- LeClercq, T., 1999. Principle 4: Good practice gives prompt feedback. J. Legal Educ. 49, 418.
- van der Lee, C., Gatt, A., van Miltenburg, E., Krahmer, E., 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. Comput Speech Lang 67, 101151.
- Leighton, J.P., Bustos Gómez, M.C., 2018. A pedagogical alliance for trust, wellbeing and the identification of errors for learning and formative assessment. Educational Psychology 38, 381–406.
- Li, C., Xing, W., 2021. Natural language generation using deep learning to support mooc learners. International Journal of Artificial Intelligence in Education 31, 186–214.
- Lim, L.A., Dawson, S., Gašević, D., Joksimović, S., Pardo, A., Fudge, A., Gentili, S., 2021. Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses. Assessment & Evaluation in Higher Education 46, 339–359.
- Lin, J., Dai, W., Lim, L.A., Tsai, Y.S., Mello, R.F., Khosravi, H., Gasevic, D., Chen, G., 2023a. Learner-centred analytics of feedback content in higher education, in: LAK23: 13th International Learning Analytics and Knowledge Conference, pp. 100–110.
- Lin, J., Thomas, D.R., Han, F., Gupta, S., Tan, W., Nguyen, N.D., Koedinger,

- 1 K.R., 2023b. Using large language models to provide explanatory feedback
2 to human tutors. *arXiv preprint arXiv:2306.15498* .
- 3 MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., Huang, Z., 2022.
4 Generating diverse code explanations using the gpt-3 large language model,
5 in: *Proceedings of the 2022 ACM Conference on International Computing
6 Education Research-Volume 2*, pp. 37–39.
- 7 Markel, J.M., Opferman, S.G., Landay, J.A., Piech, C., 2023. GPTeach: In-
8 teractive TA training with GPT-based students, in: *Proceedings of the 10th
9 ACM Conference on Learning @ Scale*, Association for Computing Ma-
10 chinery, New York, NY, USA. p. 226–236. doi:10.1145/3573051.3593393.
- 11 Middleton, T., Ahmed Shafi, A., Millican, R., Templeton, S., 2023. Develop-
12 ing effective assessment feedback: Academic buoyancy and the relational
13 dimensions of feedback. *Teaching in Higher Education* 28, 118–135.
- 14 Mulliner, E., Tucker, M., 2017. Feedback on feedback practice: perceptions of
15 students and academics. *Assessment & Evaluation in Higher Education* 42,
16 266–288.
- 17 Nicol, D.J., Macfarlane-Dick, D., 2006. Formative assessment and self-
18 regulated learning: A model and seven principles of good feedback practice.
19 *Stud. High. Educ.* 31, 199–218.
- 20 Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E., 2023.
21 Capabilities of gpt-4 on medical challenge problems. *arXiv preprint
22 arXiv:2303.13375* .
- 23 Osakwe, I., Chen, G., Whitelock-Wainwright, A., Gašević, D., Pin-
24 heiro Cavalcanti, A., Ferreira Mello, R., 2022. Towards auto-
25 mated content analysis of educational feedback: A multi-language
26 study. *Computers and Education: Artificial Intelligence* 3, 100059.
27 doi:https://doi.org/10.1016/j.caeai.2022.100059.
- 28 Pankiewicz, M., Baker, R.S., 2023. Large language models (gpt) for automating
29 feedback on programming assignments. *arXiv preprint arXiv:2307.00150* .
- 30 Pardo, A., Bartimote, K., Shum, S.B., Dawson, S., Gao, J., Gašević, D., Le-
31 ichtweis, S., Liu, D., Martínez-Maldonado, R., Mirriahi, N., et al., 2018.
32 Ontask: Delivering data-informed, personalized learning support actions.
33 *Journal of Learning Analytics* 5, 235–249.
- 34 Patel, N., Nagpal, P., Shah, T., Sharma, A., Malvi, S., Lomas, D., 2023. Improv-
35 ing mathematics assessment readability: Do large language models help?
36 *Journal of Computer Assisted Learning* 39, 804–822.
- 37 Ryan, T., Henderson, M., Ryan, K., Kennedy, G., 2023. Identifying the compo-
38 nents of effective learner-centred feedback information. *Teaching in Higher
39 Education* 28, 1565–1582.
- 40 Shute, V.J., 2008. Focus on formative feedback. *Review of educational research*
41 78, 153–189.
- 42 Weaver, M.R., 2006. Do students value feedback? student perceptions of tutors’
43 written responses. *Assessment & Evaluation in Higher Education* 31, 379–
44 394.
- 45 Yan, L., Martínez-Maldonado, R., Gašević, D., 2024. Generative artificial intel-
46 ligence in learning analytics: Contextualising opportunities and challenges
47 through the learning analytics cycle, in: *Proceedings of the 14th Interna-
48 tional Conference on Learning Analytics and Knowledge*, Association for
49 Computing Machinery, New York, NY, USA. p. in press.
- 50 Yan, L., Sha, L., Zhao, L., Li, Y., Martínez-Maldonado, R., Chen, G., Li, X.,
51 Jin, Y., Gašević, D., 2023. Practical and ethical challenges of large lan-
52 guage models in education: A systematic scoping review. *British Journal of
53 Educational Technology* .
- 54 Yang, M., Carless, D., 2013. The feedback triangle and the enhancement of
55 dialogic feedback processes. *Teaching in Higher Education* 18, 285–297.