The effects of an AI feedback coach on students' peer feedback quality, composition, and feedback experience

Rasmus R. Hansen, Aarhus University

Rikke Frøhlich Hougaard, Aarhus University

Karen Louise Møller, Aarhus University

Annika Büchert Lindberg, Aarhus University

Tobias Alsted Nielsen, Aarhus University

Christopher Neil Prilop, Aarhus University

Abstract

This study examines the integration of an Artificial Intelligence (AI) feedback coach in a peer feedback activity. Participants provided peers with feedback on their assignments. While providing feedback, they either received real-time adaptive AI coaching (intervention group) or not (control group). Feedback comments from participants were analysed concerning content, text complexity, and sentiment. Survey responses were coded for sentiment and themes. Results show adverse effects of the AI feedback coach. Intervention group participants' feedback included fewer reflective questions and adhered less to criteria. They provided shorter, more complex feedback. Students indicated mixed views on the AI feedback coach, with some finding it helpful and others distracting. A notable subset of students stated overreliance on the AI coach, prioritising its validation over their own judgment. Results suggest that AI tools need thoughtful integration, possibly with additional scaffolding to counteract overreliance and avoid negative impact on peer feedback quality and feedback experience.

ISSN: 1903-248X

Introduction

Peer feedback is a reciprocal process, during which students assess the work of a peer and provide criteria-based, constructive comments. As a highly effective teaching intervention, it has developed into a common phase in higher education classrooms (Nicol et al., 2014). Taking advantage of the independence of time and place, digital learning environments are being used to create even more opportunities for peer feedback. Yet, peer feedback can be of lower quality than expert feedback (e.g., Prins et al., 2006). To further enhance the impact of peer feedback, students must be supported in developing feedback literacy (Carless & Boud, 2018).

Educational research shows that instructional design elements such as scaffolds or prompts can successfully improve peer feedback quality (e.g., Gan & Hattie, 2014). *Prompting* is the practice of facilitating feedback or other learning strategies through reflective questions and hints (Nückles et al., 2009). While analogue peer feedback activities rely on static scaffolds or prompts, for example rubrics or a fixed set of questions, digital learning environments can now provide more *adaptive* or *dynamic* scaffolding or prompting opportunities (e.g., Nguyen, 2023; Topping, 2021). Especially since Large Language Models (LLMs) have become widely accessible and applicable for educational purposes, the potential of using Artificial Intelligence (AI) to support students' learning processes is being explored. AI can, for example, be used to provide students with scaffolds or prompts that adapt to their inputs in real time (Kasneci et al., 2023). However, while researchers acknowledge the potential of AI to enhance education, they also recognize its capacity to negatively affect student learning (Jensen et al., 2023).

This study therefore investigated the implementation of an AI feedback coach into a peer feedback activity in higher education. In a quasi-experimental design, students' peer feedback quality and their perception of the peer feedback process when taking part in a digital peer feedback activity were analysed. In the intervention group, participants' peer feedback was scaffolded and prompted by adaptive, real-time AI coaching, while participants in the control group did not receive any AI coaching. The study aimed to analyse how an AI feedback coach impacted students' peer feedback quality and their perception of the peer feedback process.

Consequently, this study sheds light on adaptive, real-time prompting and scaffolding in peer feedback activities. It provides perspectives on which impact AI can have on students' peer feedback quality and how students perceive the AI-enhanced feedback process. From a methodological standpoint, this study provides first insights into how Natural Language Processing (NLP) can assist qualitative and quantitative-qualitative data analysis in feedback research.

Theory and prior studies

Feedback

Feedback in teaching is defined by encounters, where the learner receives information about their performance, which can be formative in the sense, that it is meant to improve their learning, or modify their behaviour (Shute, 2008; Jensen et al., 2023). Feedback has been shown to be an effective tool for enhancing learning and performance in higher education (Hattie, 2023; Wisniewski et al., 2020). In recent years, feedback has been conceptualized as a co-constructive process, with a shared responsibility and partnership between the feedback provider and recipient, fostering students' self-regulated learning competencies (Carless, 2020; Nicol & MacFarlane-Dick, 2006). This conception of feedback aligns with the social constructivist learning theories, emphasising interaction as a means to construct knowledge and learning (Vygotsky, 1962). As the new feedback paradigm, which is more complex for the student than the old transmission-focused paradigm, becomes more present in higher education and is recognized as a valuable transferable skill, the need to train student feedback literacy grows (Carless & Boud, 2018).

Peer feedback

In higher education, peer feedback is often applied besides teacher-student feedback (Nicol et al., 2014). Peer feedback is a reciprocal process in which students assess their peers' work or effort and provide them with constructive comments that include possible strengths, weaknesses, and alternatives to stimulate reflection (Nicol et al., 2014; Prilop & Weber, 2023). As students both provide peers with feedback and receive feedback from their peers, students not only learn from receiving feedback comments but also develop evaluative judgement and feedback literacy by constructing comments for their peers (Carless & Boud, 2018; Nicol & Macfarlane-Dick, 2006). Contrary to expert feedback, students take an active role in peer feedback, which gives them ownership of their learning process, which can result in higher motivation and engagement (Liu & Carless, 2006; Ozogul & Sullivan, 2009).

To increase the opportunities for peer feedback, peer feedback activities are often conducted in digital learning environments. Digital learning environments allow for asynchronous feedback enabling students to engage independently from a specific time and space. Digitally mediated feedback allows teachers to use an array of tools to enhance the peer feedback process (Carless & Boud, 2018; Prilop et al., 2019).

However, research suggests that not all students demonstrate sufficient feedback literacy to provide high-quality feedback (e.g., Chen, 2014). In different domains, systematic comparisons (Prilop et al., 2021; Prins et al., 2006) of expert and novice feedback showed that expert feedback is generally of higher quality. Comparing expert and novice feedback by using a set of established criteria (Feedback Quality Index, Prins et al., 2006), showed that experts relate more to assessment criteria, are more specific, and provide more reflective questions in their feedback. Yet, research indicates that peer feedback quality can significantly be enhanced by scaffolding the peer feedback process with rubrics, exemplars, or prompts (Carless & Boud, 2018; Gan & Hattie, 2014; Panadero & Jonsson, 2013).

Peer feedback quality

Different factors in the feedback process can impact and influence the perception and reception of feedback, such as the valence of the feedback or the trustworthiness of the feedback provided, and lead to substantive, critical reflection on the part of the feedback recipient (Jensen et al., 2023; Prilop et al., 2021; Ruwe & Mayweg-Paus, 2023). Therefore, prior research has mostly focused on a set of criteria to establish peer feedback quality or its linguistic composition (e.g., Prins et al., 2006; Ryan et al., 2022). Concerning feedback criteria, studies emphasise that high-quality feedback, for example, should be based on assessment criteria, highlight specific examples, and explain evaluations so that the feedback recipient trusts and can act on the feedback. Furthermore, feedback should incorporate suggestions for inclusion and questions to stimulate reflection and engagement. Regarding linguistic characteristics, feedback length and valence have been analysed. For example, the sentiment or valence of feedback can have an impact on students' self-efficacy (Prilop et al., 2021). Concerning the length and complexity of the written feedback, both feedback that is short and vague as well as overloaded and long, can have a negative impact on the interpretation of the feedback. Feedback that is clear and simply constructed is interpreted more effectively than complexly written feedback (Ryan et al., 2022). However, as the length of feedback often coincides with the presence of other quality criteria (e.g., questions or suggestions), feedback research generally assumes elaborate feedback has a greater impact on students' learning outcomes (e.g., Van der Kleij et al., 2015).

Enhancing the peer feedback process

Prompting and scaffolding peer feedback

Learning processes can be supported with scaffolds and prompts. Scaffolds or prompts are commonly used in teaching and learning, as questions or hints that are meant to cause reflection and induce productive learning strategies (Nückles et al., 2009). Scaffolds and prompts guide the learner through a task and are often defined by drawing students' attention to the important part of the process while fading, partially leaving the learner to problem-solve by themselves (Van de Pol et al., 2010; Wood et al., 1976). Scaffolds and prompts can lessen learners' cognitive load (Nückles et al. 2009). According to cognitive load theory, learning experiences can cause extraneous, intrinsic, and germane cognitive load for learners (Sweller, 1994). While intrinsic cognitive load is the inherent complexity of a task, which cannot be reduced, extraneous cognitive load is caused by the way the task is presented. The presentation can result in unnecessary difficulties if designed poorly. The goal of every learning task should be to increase germane cognitive load as this is the load induced by actually processing information, i.e., learning. Hence, if designed appropriately, scaffolds and prompts can make it easier for learners to understand a task (decrease extraneous cognitive load) and engage with the learning task (increase germane cognitive load).

Prompting can also promote students' peer feedback quality. Feedback prompts are then questions or sentence openers meant to guide feedback providers, through cues and suggestions, to focus the feedback on specific elements, to improve the overall quality of the feedback (Gan & Hattie, 2014). Studies show that using question prompts to guide students when providing peer feedback can help guide the feedback to focus on learning gaps and ways to improve performance (Gan & Hattie, 2014). Scaffolding peer feedback activities have also been found to increase the quality of feedback specifically helping students provide better self-regulation feedback, but only for students of medium or high domain knowledge (Alqassab et al., 2018). In other domains, such as writing reflective learning protocols, prompts have been found to stimulate cognitive and metacognitive strategies, which also indicates that prompts can be designed to support specific relevant teaching and learning strategies (Nückles et al., 2009). However, studies (Nückles et al., 2010) show that prompts can also have detrimental effects on learners, when prompts go from being a helpful tool to activating cognitive and meta-cognitive strategies, an *expertise reversal effect* can take effect if the students have internalised the prompted strategies, making the prompts instead a redundant stimulus, inducing extraneous cognitive load.

Artificial Intelligence in the Peer Feedback Process

Though the implementation of AI in education is not new, the release of ChatGPT in November 2022 has introduced GenAI tools at scale to the education sector. This has led educators and educational developers to speculate and point to the possible ways the educational field can and will be impacted by the developments in this technology (Kasneci et al., 2023; Malmström et al., 2023). The generative and humanlike nature of GenAI offers new opportunities and different challenges when applied in an educational context, for example, that the rising number of AI tools in education could lead to students over-relying on them (Darvishi et al., 2024).

Earlier research concerning AI in feedback processes studied automated feedback, and thus some principles of pre-2022 research into AI in education obviously remain relevant (Zhai et al., 2021). Since the public launch of ChatGPT, the addition of LLM-generated feedback on written products has been found to positively influence students' revision performance, motivation, and evoke positive emotions (Dell'Acqua et al., 2023; Meyer et al., 2024). In a study by Ruwe and Mayweg-Paus (2023), the participants assessed input as more trustworthy when they were told it originated from an AI model. Furthermore, Jansen et al. (2024) found that LLM-generated feedback was perceived as useful, not on the level of expert feedback, but sufficient to give to students in most instances. However, research shows that the trustworthiness and the perceived competence level of the feedback provider can also render students passive. They rely too much on the judgment of the feedback provider instead of reflecting for themselves (Strijbos et al., 2010). Furthermore, technology acting as an agent in the feedback process (Panadero & Lipnevich, 2022), can lead to unpredictable behaviour when systems utilise GenAI with little human supervision (Kasneci et al., 2023).

Yet, by utilising technology (Deeva et al., 2021) that can react in real-time to what students are writing, and update as the writing changes, adaptive prompting has become possible. Personalised or adaptive prompting could further strengthen the usefulness of prompts. This interactivity of elements in a learning process has been shown to positively affect, for example, motivation, and can thus lead to enhanced performance and an increase in the depth of learning and understanding (Evans & Gibbons, 2007; Kettanurak et al., 2001). In other domains, adaptive prompting has been examined concerning prompts adapting to, for example, students' knowledge level, which had a positive impact on students' learning compared to students receiving non-adaptive prompting (Bimba et al., 2021). Moreover, a study into the adaptive prompting of pedagogical agents on student learning and self-regulation found that the quality and not the frequency of the adaptive prompting had a positive impact on the self-regulated learning competencies of students (Bouchet et al., 2016).

These effects can be aligned with Mayer's cognitive theory of multimedia learning (2002). Building on cognitive load theory, Mayer's theory focuses on how multimedia elements in learning environments can inflict positive or negative cognitive load. Though the interactivity of adaptive prompts can be assumed to draw students' attention to relevant information (reducing extraneous cognitive load; increasing germane cognitive load), a risk of distraction also follows with the use of interactive multimedia elements (Harp & Mayer, 1998). When interactive elements become distracting, students' ability to focus their attention on a given task is harmed as extraneous cognitive load increases due to higher complexity (Lawson & Mayer, 2024).

Research questions

High-quality peer feedback can significantly enhance students' learning outcomes. However, low peer feedback quality can harm the peer feedback process. Therefore, this study applied a quasi-experimental design to investigate how adaptive, real-time scaffolds and prompts by an AI feedback coach can impact students' peer feedback quality and perception of the peer feedback process.

Specifically, the study addressed the following research questions (RQ):

RQ1: How does an AI feedback coach impact students' peer feedback quality?

RQ2: How does an AI feedback coach impact the composition and sentiment of students' peer feedback comments?

RQ3: How do students experience being coached by an AI feedback coach, while providing written peer feedback?

Methods

Study context

The study was conducted in a pedagogical course (3 ECTS) for PhD students who work as teaching assistants (TAs) within STEM disciplines at a Danish university. The course was designed as a 16-week blended-learning format including two initial weeks of online learning, three on-campus course days, and a project period in between the on-campus course days 2 and 3. All teaching and exercises were in English.

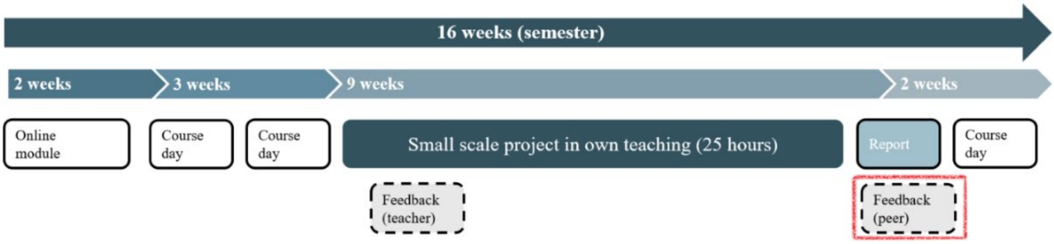


Figure 1: Timeline of the PhD course. The peer feedback activity is shown within the final 2 weeks, marked with a red outline.

The assignment of the course is a project report (4-8 pages) on a project that the participants plan in their own teaching. This assignment is handed in three times: initial, draft, and final (see the “perforated” boxes in Figure 1 - the second highlighted box represents both the 2nd and 3rd hand-in). With the initial hand-in, the students receive teacher feedback on the design of the teaching project and for the draft hand-in they receive peer feedback on their project report, using the LMS extension FeedbackFruits, and an included rubric. The rubric is the final part of the peer assessment, where some criteria (e.g. evaluating the report’s description of “Problem and Aim” and its “Planned initiative”) let the students provide more structured feedback, compared to the more open initial comments. The final version of the plan only receives feedback in the form of the teachers’ approval of their hand-in. The project period spans 9-12 weeks. The course is designed on the foundation of the model for teacher professional growth to promote the TAs professional growth by facilitating professional discussions about shared experiences (Clarke & Hollingsworth, 2002; Nielsen, 2012).

Sample

Data were collected from PhD students enrolled in two iterations of the course. In the first iteration ($_1$), a quasi-experimental study was conducted to compare an intervention group (IG$_1$) with an enabled AI feedback coach with a control group (CG$_1$) without an AI feedback coach. In total, 46 students consented to participate in the study, with 20 students (38% female, 62% male, 0% non-binary, 0% undisclosed; M_{Age}: 27.61) in the IG$_1$ and 26 students (43% female, 53% male, 0% non-binary, 3% undisclosed; M_{Age}: 27.87) in the CG$_1$. [i] To control between groups, the participants were asked the question “How would you assess your peer feedback competence concerning teaching?” to assess their own feedback competence on a 5-point Likert scale from “very low” to “very high”. No significant differences were found (a*verage self-assessed peer feedback competence:* CG = 3.3; IG$_1$ = 3.2).

In the second iteration of the course ($_2$), the group of PhD students took part in the peer feedback activity with the AI feedback coach enabled. 62 students (40% female, 57% male, 2% non-binary, 2% undisclosed; *average self-assessed peer feedback competence*: 3.4, M_{Age}: 27.47) gave consent.

Distinct data were collected. In the first iteration, students’ feedback comments were analysed and compared between groups. In the second iteration, student answers concerning their experience with the AI feedback coach were collected.

Procedure

During the competency development course, students were given an assignment that consisted of them writing an individual project report about a project in their teaching. After uploading the project report to the digital feedback environment (see ”The AI feedback coach”), each student was tasked with providing peer feedback to two of their peers. Consequently, all students also received feedback from two of their peers.

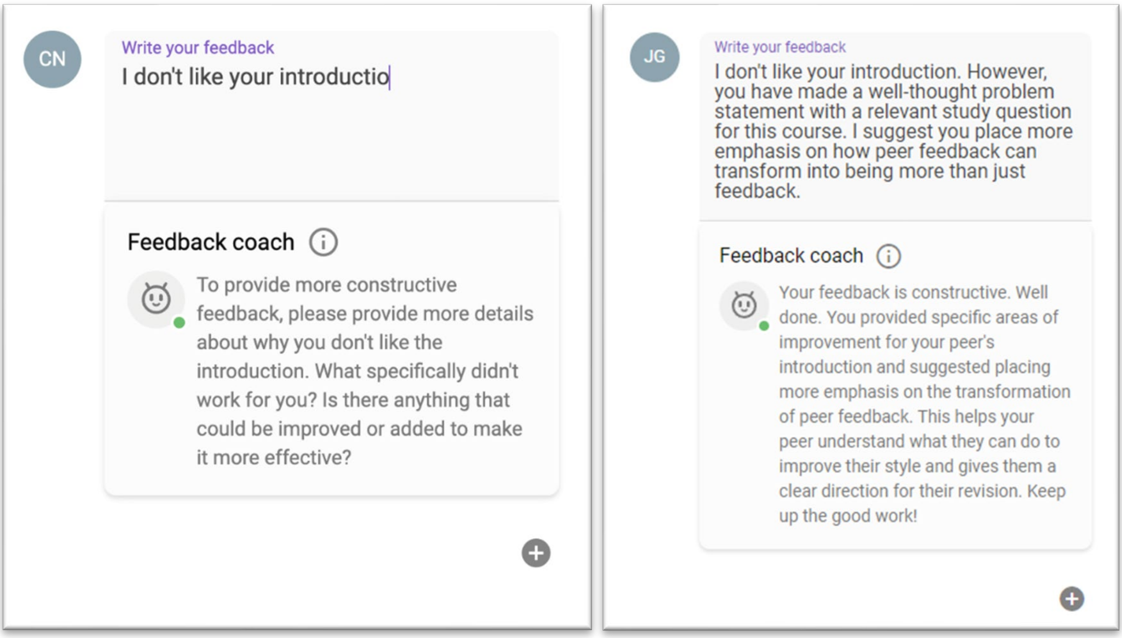
In the first iteration, an intervention group had access to the AI feedback coach enabled in the digital feedback environment (see ”The AI feedback coach”), while the control group had the same task description and digital environment, except for the lack of the AI feedback coach. In the second iteration, participants took part in the same intervention as the IG of the first iteration (AI feedback coach enabled).

All participants used the digital peer feedback platform FeedbackFruits within the university's LMS. When a student has uploaded their project report, the feedback environment assigns them two peers to whom they must anonymously provide feedback. These peers are randomly selected. For each project report they provide feedback on, they must submit at least four feedback comments about what parts are good and what needs improvement. In the feedback environment, they are presented with a peer's project report, in which they can highlight certain passages to connect their individual comments. After submitting their comments, they are presented with a rubric to give an overall assessment of the project report within predefined criteria. After two peers have reviewed a project report, the author of the report is given access to the feedback and assessment. The exercise ends with a reflective phase, where the student is asked to consider and write down how the received feedback helped them, and what knowledge and inspiration could be drawn from the process and providing feedback on the other project reports.

The AI feedback coach

The digital learning environment, developed by the company FeedbackFruits, allows an AI-driven feedback coach to be enabled. The AI feedback coach delivers adaptive meta-feedback on students' feedback comments by utilizing the LLMs of Azure's OpenAI services (Nguyen, 2023)[ii]. FeedbackFruits (2024) discloses that the AI feedback coach has explicitly been instructed to “use good feedback practices to ensure constructive responses to students' comments.”. While the exact prompts are not disclosed to the public, FeedbackFruits highlights six key areas of good feedback practice that the AI has been primed with (FeedbackFruits, 2024):

- Respond in a helpful and coachable way.
- Ensure that the feedback addresses the given criteria.
- Include the feedback and context (such as the criteria) in a specific format.
- Compliment and encourage students to give better feedback.
- Do not respond to gibberish, random text, or things that appear unrelated to giving feedback to a peer.
- Offer areas of improvement.



I don't like your introductio

To provide more constructive feedback, please provide more details about why you don't like the introduction. What specifically didn't work for you? Is there anything that could be improved or added to make it more effective?

I don't like your introduction. However, you have made a well-thought problem statement with a relevant study question for this course. I suggest you place more emphasis on how peer feedback can transform into being more than just feedback.

Your feedback is constructive. Well done. You provided specific areas of improvement for your peer's introduction and suggested placing more emphasis on the transformation of peer feedback. This helps your peer understand what they can do to improve their style and gives them a clear direction for their revision. Keep up the good work!

Figure 2: Two screenshots from the digital feedback environment, with the AI feedback coach enabled, showing the AI adapting its coaching to the written feedback comment.

The AI operates by tokenizing students' feedback comments sending them to Azure's OpenAI Services to be processed by an LLM. The relevant criterion or rubric is also sent along with the student's feedback comment, but not any part of the object (in this case the written project report) that the feedback is directed towards. The LLM then generates a scaffolding prompt in response (see Figure 2). This response is sent back to the platform where it appears directly below the student's feedback comment in real time (FeedbackFruits, 2024).

Measures

The data from the quasi-experiment in the first iteration of the course (i.e. the feedback messages) were analysed concerning feedback quality, sentiment, and linguistic metrics. The data from the second iteration (i.e. the answers to the open-ended questions) were analysed using deductive and inductive qualitative analysis. These analytical approaches are explained in the following sections.

Feedback quality

From the first iteration, participants' peer feedback comments (IG_1 and CG_1) were exported from the digital feedback environment and analysed. In total, participants wrote 396 feedback comments. The feedback comments were analysed using an adapted version of Prins et al.'s (2006) feedback quality index. The feedback quality index assesses feedback comments according to a set of criteria and evaluates whether the feedback is of 'good', 'average', or 'suboptimal' quality. The following criteria were used to code the feedback comments: Content/Criteria, Explanations/Examples, Specificity, Questions, Advice/Suggestions, First-person (Style), and Encouragement (see Table 1).

An example of a code 2 comment of the specificity category is 'Really good the procedure of you explaining, the thinking in pairs, leaving them to develop a bit more on their own and they having the chance of speaking with others'. For the advice/suggestions category, a comment such as 'I recommend that you add a couple of lines at the end to sum up you initiative, to make sure your intentions for lesson is clear for all readers.' was coded '2'.

For the quantitative-qualitative coding procedure, test codings of 10% of the total sample (40 comments), with subsequent discussions of interpretations, were conducted by two coders, to determine interrater-reliability. Every test used a new sample of a random 10% of the feedback comments and was repeated three times. The coding was done according to the coding manual and Cohen's Kappa was calculated to establish inter-rater reliability (see table 1). Kappa values ranged from substantial to almost perfect agreement for all categories.

Table 1: Content analysis of PhD students' peer feedback: category, feedback quality, and Cohen's K (Interrater reliability).

Category	Good feedback (coded as 2)	Average feedback (coded as 1)	Sub-optimal feedback (coded as 0)	Cohen's κ
Content/ Criteria	Substantial subject-specific (in this case concerning didactics) remarks	Some subject-specific related remarks (could be about formatting, writing style etc.)	No or hardly any subject-specific related remarks	0.624
Use of explanations/ Examples	Description and explanation of remarks throughout (on all remarks)	Some descriptions and explanations of remarks	No description of behaviour and no explanation of remarks	0.609
Specificity	Reference to specific parts of the project report	Reference to a section of the project report (including @-mentions)	No specific references	0.736
Questions	Activating questions fostering reflection	Clarifying questions (confusion)	No questions	0.811
Advice/ Suggestions	Good and clear suggestions for improvement; constructive advice	Some suggestions for improvement	No suggestions for improvement	0.666
Style (first person)	First-person throughout	Sometimes in the first-person	No first-person	0.776
Formulation (Encouragement)	High praise/Strong encouragement	Some encouragement	No encouragement	0.750

Metrics for length, sentiment, and readability

The same 396 feedback comments were also analysed concerning composition, readability, and sentiment. The analysis utilized the Python packages TextDescriptives and VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis[iii], which provide ready-to-use algorithms for calculating the relevant descriptive values based on the supplied text input. These packages were chosen and included to calculate some of the linguistic characteristics that have been known to affect feedback processes and quality, namely complexity, length, and valence (see "Peer feedback quality").

The TextDescriptives library utilises a spaCy pipeline and makes it possible to calculate different metrics from text, as the package utilises LLMs, such as BERT to complete tasks such as tokenization, sentence segmentation, part-of-speech tagging and more, based on user text input (Hansen et al., 2023; Honnibal et al., 2020). TextDescriptives allow for the calculation of, among other things, different measures of text complexity and readability, such as 'Lix', 'Rix', 'Coleman liau', and 'Flesch Reading Ease'. All of these are measures of text complexity or readability calculated with pre-defined algorithms using variables such as syllables, characters, tokens, and words per sentence and/or the entire text.

VADER sentiment analysis is a commonly used, freely available, and well-assessed Python Library, that allows the user access to a lexicon and rules-based model for calculating sentiment, where a lexicon of sentiment-weighted words are combined with grammatical and syntactical rules to output a number between -1 and 1, describing to what degree a text is negative (below zero) or positive (above zero) (Hutto & Gilbert, 2014; Ribeiro et al., 2016). Hutto & Gilbert (2014) benchmarked and compared VADER and other sentiment analysis methods, finding that it outperforms human annotators and other machine learning models made for the same task. An ad-hoc quality check applied to a small sample of our data supported its high reliability.

The GitHub links to both TextDescriptives and vaderSentiment can be found in the appendices.

AI feedback coach perception

In the second iteration of the quasi-experiment, the participants answered a short questionnaire the week after they had completed the peer feedback task.

The questionnaire established demographic information but otherwise focused on the two questions:

- How did you experience receiving ‘coaching’ from the AI feedback coach?
- How did the AI coach impact the way you wrote your feedback comments?

Data analysis

To compare the characteristics - linguistic metrics, and the quality of the comments between the groups - statistical analysis, in the form of independent *t*-tests, was conducted. Statistical calculations were performed using Python and SPSS 29.

The answers to the open-ended survey questions were initially coded deductively to ascertain the students’ overall sentiment towards the AI feedback coach. A scale of positive, somewhat positive, neutral, somewhat negative, and negative was used for this. Two of the authors coded the answers blindly, and then compared and discussed inconsistencies until agreement was found on all coded answers (Zottmann et al., 2013).

Then the answers were all inductively thematically analysed through coding by the team of authors (Braun & Clarke, 2006). The coders initially focused on each individual comment and ascertained implied meaning and reasoning. By iteratively close-reading comments, common themes emerged and could be validated and defined through discussion and categorisation in the research group. Finally, each theme was supplemented with quotes from several students' answers, thereby giving depth, motivation, and nuance to the overarching themes. This analysis was done in order to explore themes regarding how the students perceived the impact of the AI feedback coach on the peer feedback process and with the intent of uncovering underlying reasons and explanations for any differences in feedback composition and quality.

In this way, the study takes on a mixed methods approach, where qualitative and quantitative data and methods combine to complement each other, and where one method can enlighten the findings of the other through their differences (Creswell, 2014).

Results

In the following section, the results from the three different analyses are presented including a short interpretation of each. First, the coding of the feedback quality index is presented, then the metrics concerning text composition, readability, and sentiment, and lastly the results from the qualitative analysis of the survey answers are presented.

RQ1: Feedback quality

The analyses of differences in feedback quality between participants of the CG and IG showed that members of the IG included significantly fewer reflective questions, $t(44)= 2.155$, $p=.031$, $d=.217$, and their feedback was significantly less criteria-based, $t(44)= 2.014$, $p=.044$, $d=0.203$ (see Table 2). Though Cohen's *d*-value of 0.2 only suggests a small effect, the results still indicate that the influence of AI coaching could significantly impact what type of feedback students provide (Cohen, 1988). Analyses of the other feedback quality categories revealed no significant differences, $p>.05$.

Table 2: Means, standard deviations and differences between CG and IG for each category of the feedback quality index.

	CG		IG		
	M	*SD*	*M*	*SD*	Δ
Content/Criteria	1.44	0.71	1.30	0.74	-0.15
Use of explanations/ Examples	1.04	0.95	1.01	0.96	-0.03
Specificity	0.81	0.65	0.76	0.62	-0.05
Questions	0.18	0.52	0.09	0.35	-0.10
Advice/Suggestions	0.54	0.77	0.52	0.73	-0.02
Style (first person)	0.62	0.83	0.59	0.80	-0.03
Formulation (Encouragement)	0.76	0.86	0.71	0.82	-0.05

RQ2: Length, sentiment, and readability

Analyses showed no significant differences between the CG and IG concerning the sentiment score. Both means were found to be positive (M_{IG}=0.460, M_{CG}=0.467). This indicates that the AI coach did not significantly influence the valence or amount of positive or negative wording, such as praising or berating language, in the feedback comments, and that the students in both groups were on average more positive than negative when giving feedback.

Concerning the metrics calculated using the TextDescriptives library, several significant differences were found between the two groups (see Table 3). These differences concerned the length, and composition, as well as the readability of the comments.

Table 3: Means, standard deviation, and delta-values for significant metrics concerning text composition and complexity.

	CG		IG		
	M	*SD*	*M*	*SD*	Δ
Text composition					
Sentence length	19.46	6.80	15.45	4.30	-4.01
Tokens per comment	74.27	35.08	58.93	26.48	-15.34
Syllables per token	0.68	0.10	0.82	0.18	0.14
Text complexity					
Flesch Reading Ease	73.29	9.64	61.84	19.94	-11.45
Coleman Liau	8.52	2.05	11.88	4.43	3.36
Lix	40.30	8.87	46.41	13.07	6.11
SMOG	9.72	1.98	10.97	2.63	1.25

Note: For most complexity metrics, higher values indicate greater complexity. However, for the Flesch Reading Ease metric, a higher score denotes easier readability, meaning lower complexity.

IG members used significantly shorter sentences, *t*(44)=3.274, *p*=.001, *d*=0.7, and used fewer tokens per comment, *t*(44)= 2.285, *p*=.024, *d*=0.491. Furthermore, IG members had a significantly higher count of syllables per token, *t*(44=-4.969, *p*<0.001, *d*=1.108). Finally, members of the IG wrote significantly more complex feedback comments, as measured by several measures for text complexity, that uses metrics such as sentence length and syllables per word to calculate complexity. These metrics are ‘Flesch Reading Ease’, *t*(44)=3.331, *p*=.001, *d*=0.739, ‘SMOG’, *t*(44)=-2.459, *p*=.016, *d*=0.539, ‘Coleman Liau’, *t*(44)=-4.437, *p*<.001, *d*=0.985, and ‘Lix’, *t*(44)=-2.505, *p*=.014, *d*=0.551.

The effect sizes for these calculations were all medium to large, with the majority being of large effect. This indicates that the AI feedback coach had a large and significant effect on the textual composition of the peer feedback comments and that members of the IG wrote shorter but more complex comments.

RQ3: Students' perception of and experience with the AI feedback coach

Overall sentiment toward the AI feedback coach

The following table shows the distribution of the qualitatively assessed overall sentiment of the students' answers to the question “How did you experience receiving ‘coaching’ from the AI feedback coach?”:

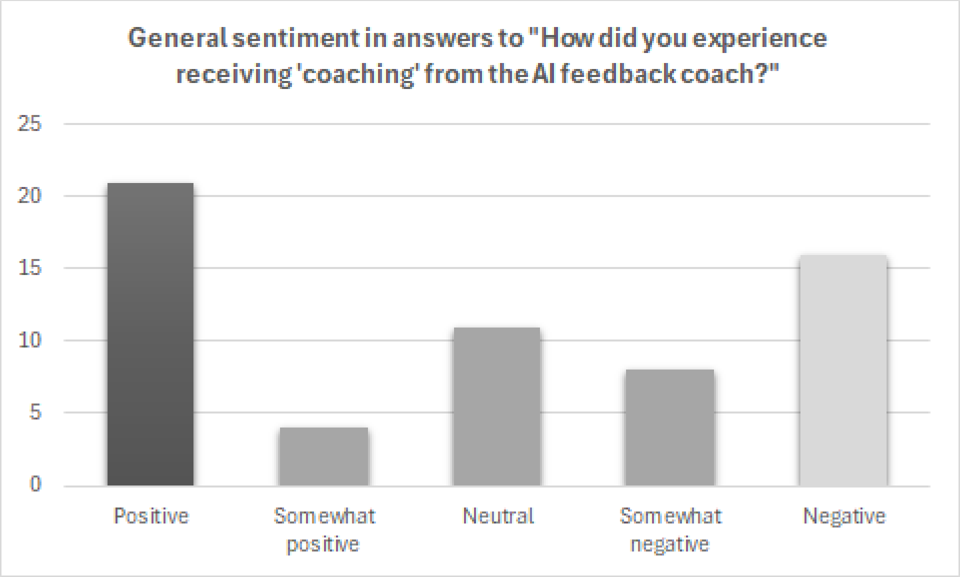


Figure 3: The distribution of sentiment in participants' descriptions of their experience with the AI feedback coach.

A total of 25 answers were coded as ‘positive’ or ‘somewhat positive’, while 11 were neutral, and 24 were ‘negative’ or ‘somewhat negative’ (see Figure 3). This showcases that a nuanced understanding of the student experience of working with the AI feedback coach is needed.

Qualitative analysis of students’ comments

The thematic analysis of students’ answers to the open-ended questions revealed five distinct themes: Validation from the AI, Appeasing the AI, Motivation, Focus on valence, and Distraction & Annoyance. The themes shed light on the almost equal distribution of sentiment towards the AI feedback coach.

Validation from the AI

> “First I thought it was nice to get confirmation that what I wrote was good feedback, but I didn't think about the suggestions it gave.”

> “It was nice to see the AI feedback coach give me feedback on my feedback. This was a nice assurance that my feedback may actually be valuable.”

Among the most prevalent recurring themes were students expressing feeling validated by the AI’s coaching. The feeling of validation seems to have assured them that their feedback was valuable and useful, and they used the AI’s coaching to support their self-assessment. Students explained being prompted to reflect on their own feedback practice, and that even if they at times did not agree with the AI’s assessment, they still enjoyed having a second opinion to rely on.

Appeasing the AI

> “I don't think [about] if my suggestions are good, I merely write until the AI accepts it.”

> “Sometimes I stopped writing when the AI already showed "Good Feedback", so I ended my point but didn't [write] much more.”

> “I spent more time trying to optimise my responses to get a good "grade" from the bot.”

On the other end of the spectrum, some students expressed a desire to simply write feedback in a way that appeased the AI coach. Some students explicitly mentioned stopping writing feedback as soon as the AI gave them a positive assessment, and perhaps sometimes stopping short or not necessarily being reflective about how to best formulate their feedback. These students also seem to disconnect the feedback exercise from the learning opportunity that this presents to both them and the feedback recipients. Several participants explained that the validation of the AI impacted their writing in an unintended way: they chose to delegate the evaluation of their feedback entirely to the AI and stopped thinking about whether the feedback they were writing was good for the intended task, but instead wrote and changed their comment until the AI was “satisfied” – not reflecting or assessing the quality of their comments themselves.

Motivation

> “Without the AI tool I think I would usually stop writing a sentence or two earlier, but this provided me the motivation to maybe push it a bit further.”

> “It forced me to be more thorough with my feedback. I would have otherwise skimmed over details about the feedback (and specifically not brought up as many examples).”

Students explained that the AI motivated them to keep writing and providing better feedback. Some of the qualities that they mentioned the AI pushed or encouraged them to be more aware of are also present in the feedback quality index, such as being specific and using explanations and examples (Prins et al., 2006). These students also seem to have been more aware of the recipient and attentive to the usefulness of their comments.

Focus on valence

> “It helped to make [the comments] more constructive and wholesome. It also showed that some comments I had written naturally had the potential of sounding mean or unhelpful, even though my intention was not that.”

> “I think it was nice to get feedback on my wording, to make my feedback as constructive and concise as possible. Specifically wording that could be misunderstood or interpreted as negative. “

> “It helped me elaborating on my feedback, perhaps allowing the receiver to get a better understanding of my feedback."

Some students mentioned that the AI helped them write their comments more constructively or positively, mentioning that the AI would help them recognise when their comment could be perceived negatively by the recipient. On the other hand, some students mentioned this attention to positivity as a downside, as they felt pushed to include positive phrasings even when they felt their peers’ product did not allow for a positive or constructive comment.

Distraction & Annoyance

> “I found it slightly annoying as it continuously updated while you were writing…”

> “I personally did not like it since the AI gave me feedback before I was done writing the comment. […] I did not feel like I needed the AI to coach me because I know how to give constructive feedback and I did not enjoy the experience of an AI telling me what to do when I did not personally ask it to.“

> “I was annoyed that it wanted me to be more positive when there was nothing to be positive about.”

Apart from feeling unnecessarily forced to construct positive feedback comments, participants stated that they were distracted by or annoyed at the AI coach, which negatively affected their ability to provide feedback. Much of the annoyance focused on the continuous updating of the

AI's suggestions being distracting. The AI assessing their feedback comment before they felt they had finished writing it was detrimental to the feedback process. These types of comments often depict the AI coach as an obstacle to the assignment, making it more difficult to focus on the task at hand.

Discussion

This investigation into a higher education peer feedback activity showed that the integration of an AI feedback coach had a significant impact on students' peer feedback quality, composition, and students' experiences of the peer feedback process. Members of the IG wrote significantly shorter and more complex feedback comments and were less likely to utilise questions for reflection and adhere to the criteria of the feedback exercise. The overall sentiment of the feedback did not differ significantly between the two groups.

Thematic analysis showed that some students became motivated by the AI's coaching and felt validated in the usefulness of their feedback comments when the AI gave them a positive assessment. Some also became more aware of the perceived valence of their comments and, therefore, reflected on the wording that they used. Meanwhile, other students found the AI detracting from the exercise of writing peer feedback, as they were distracted and became annoyed at the constant judgement and input. In contrast to the students feeling validated by the AI, a group of students instead reacted to the AI's assessment by trying to appease it.

Concerning the quality of peer feedback, intervention group members' peer feedback was significantly lower in quality regarding reflective questions and assessment criteria. This finding contradicts other studies that found prompts had a positive impact on students' peer feedback quality or other writing processes (e.g., Gan & Hattie, 2014; Nückles et al., 2009). Results from the thematic analysis indicate possible reasons for the detrimental effect. Some students seemed to perceive the AI feedback coach as an authority, taking its input as direct validation of their feedback. This concurs with findings from Strijbos et al. (2010) or Ruwe and Mayweg-Paus (2023) on the trust placed in highly competent peers or AI by students. In Strijbos et al.'s study, feedback by a highly competent peer seemed to render the students passive. Hence, it can be assumed that the AI feedback coach had a similar effect on students, causing overreliance, and that students did not display *epistemic agency*, which is their ability to actively and autonomously (in contrast to passively) engage in knowledge construction (Dai et al., 2023; Darvishi et al., 2024).

The fact that specifically reflective questions and the criteria categories were the categories where IG members' peer feedback quality significantly diverged from CG members', seems to be an effect of the specific AI feedback coach model. The AI feedback coach does not take the content of the feedback into account but only monitors surface features, hence, adherence to specific criteria is very limited. Furthermore, posing questions is not included in FeedbackFruits' key areas of good feedback. Although the exact prompts used for the AI feedback coach are not disclosed to the public, the results indicate it did not foster the provision of questions. In combination with students wanting validation from the AI feedback coach or simply wanting to appease it, this can be assumed to result in fewer questions than in the feedback provided by students in the unprompted CG. This indicates that criteria adherence and reflective questions are some of the harder feedback quality characteristics to grasp, as they were also among the characteristics found to differ between expert and novice feedback. As the participants can be considered novice feedback providers, the AI coach influence can be thought to have the most detrimental effect on feedback quality categories where novices have shown the least competence (e.g., Prilop et al., 2021; Prins et al., 2006).

The thematic analysis revealed that some students did not perceive the peer feedback process as a co-constructive learning process (Carless & Boud, 2018) but seemingly wrote their feedback without considering their own perception of feedback quality and without having the best interest of themselves and the feedback recipient in mind, instead focusing only on appeasing the AI. This indicates that some students do not grasp the value of peer feedback and lack feedback literacy (Nicol & MacFarlane-Dick, 2006). Consequently, there is a need to foster students' feedback literacy to exploit the full potential of feedback processes (Carless & Boud, 2018).

Concerning the composition of the feedback, the intervention group wrote shorter and more complex feedback comments than the control group. It can be assumed that the adaptive, real-time prompting of the AI coach distracted the students from the actual task and that an expertise reversal effect could cause annoyance for students who felt comfortable providing feedback without AI assistance (Nückles et al., 2010). This contradicts findings showing that adaptive prompting can have a positive impact on students' performance by guiding their attention to relevant information (e.g., Bimba et al., 2021; Evans & Gibbons, 2007). It can be assumed that the AI coach incurred additional extraneous cognitive load (Sweller, 1994) by adding the task of 'appeasing the AI' to the feedback process. Nonetheless, some students also reported additional motivation from being coached by the AI. This reflects the findings of other studies in the field of interactivity and adaptive prompting (Bouchet et al., 2016; Kettanurak et al., 2001). Hence, the findings show that instructional design needs to be tailored to individual students – one size does not fit all.

Implications

AI-generated feedback has the potential to be a useful add-on where the resources for expert or peer feedback are lacking, and this study clearly shows that students' feedback writing practice was impacted by the integrated adaptive feedback of the AI coach (Meyer et al., 2024; Joyner, 2017). However, the result of this study also indicates that the implicit trust that students place in AI's objectivity can have detrimental effects (Darvishi et al., 2024; Ruwe & Mayweg-Paus, 2023). It can be assumed that AI literacy, the competencies that enable critical evaluation, communication, and collaboration with AI, can play a decisive role here (Long & Magerko, 2020). An increase in AI literacy may impact students' critical reflection on AI coaching. Furthermore, detailed instruction and scaffolding of the role of the AI coach in the activity could have mitigated these effects.

As the results of this study indicate that the introduction of an AI coach has an impact on the peer feedback process, even if the results can be interpreted as negative, it still opens an avenue for designing learning activities or feedback platforms where this effect can be harnessed for positive impact instead. By tutoring the students in applying the AI coach as a tool, students could potentially limit their distraction, confusion, and blind adherence to the AI's assessment. To gain further insight into this, future research should investigate the effects of the AI feedback coach in combination with AI literacy training or a worked-example prior to the feedback task.

As the participants were all PhD students, the negative impact of the AI feedback coach could also be based on the expertise reversal effect. Many basic principles of providing high-quality feedback might already have been known to the students. Hence, participants may not have needed coaching during the feedback process, incurring extraneous cognitive load. In that case, the AI coach's prompting becomes an unnecessary distraction instead of a helping guidance (Nückles et al., 2010).

This study also supports the sentiment in other recent studies that indicate a need for critical reflection and teacher scaffolding when AI tools are introduced into teaching practices (Darvishi et al., 2024). As university students in many cases are likely to think positively of AI and believe AI to be more trustworthy than humans, the integration of AI tools cannot work without proper scaffolding and rethinking traditional didactics (Malmström et al., 2023; Ruwe & Mayweg-Paus, 2023).

Limitations

Even if the results presented are promising in understanding the impact of AI on the feedback writing process, the study still has some limitations that must be acknowledged. As the AI is a proprietary model, and we do not have full access to the prompts and pre-prompting that the system utilised, the interactions with the AI could not be systematically controlled or recorded. Therefore, it is challenging to fully understand the potential impact of prompt phrasing on the observed outcomes. Future research should attempt to document prompting and pre-prompting designs to better understand their role in influencing the process.

Though the course had over 60 students in each group, the sample size of the final groups was unfortunately quite low, as less than half of the students consented to have their feedback analysed. This means that the smaller sample size could impact generalizability. Future studies with larger or more diverse samples could help confirm these findings and explore possible variability in different samples.

The use of two parallel groups in a competency development course meant that the research design was quasi-experimental, and ecological validity was therefore prioritized over randomization. The design means that causality cannot be firmly established, but as a trade-off, the findings are likely more applicable to typical educational environments.

Unfortunately, the study design did not include a pre and post-test, which means that a development in subjects' learning and attitudes could not be analysed in depth. Future studies could introduce a pre- and post-survey with questions regarding subjects such as feedback literacy and perception, as well as AI literacy, in an attempt to further understand how the interactions with the AI impacted the beliefs and competencies of the participants.

Conclusion

This study contributes to the understanding of AI-based tools in higher education and specifically how an AI feedback coach impacts students' peer feedback provision and experience. The findings show that while the AI feedback coach influenced students' feedback process, effects were mostly negative. These results highlight the importance of tailoring AI-based tools to students' needs and emphasize the need for further research to optimize AI tools in higher education.

References

Alqassab, M., Strijbos, J. W. & Ufer, S. (2018). Training peer-feedback skills on geometric construction tasks: role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education, 33*, 11–30. https://doi.org/10.1007/s10212-017-0342-0

Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Ibrahim, S. U., Mustafa, N., Supa'at, I., ..., & Ahmad, M. Y. (2021). The effects of adaptive feedback on student's learning gains. *International Journal of Advanced Computer Science and Applications, 12*(7), 68-80. https://doi.org/10.14569/IJACSA.2021.0120709

Bouchet, F., Harley, J.M., & Azevedo, R. (2016). Can adaptive pedagogical agents' prompting strategies improve students' learning and self-regulation? In A. Micarelli, J. Stamper, & K. Panourgia (Eds). *Intelligent Tutoring Systems (Lecture Notes in Computer Science, 9684*, 368–374). Springer, Cham. https://doi.org/10.1007/978-3-319-39583-8_43

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

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

Chen, T. (2014). Technology-supported peer feedback in ESL/EFL writing classes: a research synthesis. *Computer Assisted Language Learning, 29*(2), 365–397. https://doi.org/10.1080/09588221.2014.960942

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*(8), 947–967. https://doi.org/10.1016/S0742-051X(02)00053-7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches* (4th ed.). Thousand Oaks, CA: Sage.

Dai, Y., Liu, A., & Lim, C. P. (2023). Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education. *Procedia CIRP, 119*, 84–90. https://doi.org/10.1016/j.procir.2023.05.002

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

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education 162*, 104094. https://doi.org/10.1016/j.compedu.2020.104094

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper,* (24-013). https://doi.org/10.2139/ssrn.4573321

Evans, C., & Gibbons, N. J. (2007). The interactivity effect in multimedia learning. *Computers & Education, 49*(4), 1147–1160. https://doi.org/10.1016/j.compedu.2006.01.008

FeedbackFruits (Accessed August 2024). Acai Coach Transparency Note. Retrieved from https://help.feedbackfruits.com/en/articles/7314625-acai-coach-transparency-note

Gan, M. J. S., & Hattie, J. (2014). Prompting secondary students' use of criteria, feedback specificity and feedback levels during an investigative task. *Instructional Science, 42*(6), 861–878. https://doi.org/10.1007/s11251-014-9319-4

Hansen, L., Olsen, L. R., & Enevoldsen, K. (2023). TextDescriptives: A Python package for calculating a large variety of metrics from text. *Journal of Open Source Software, 8*(84), 5153. https://doi.org/10.21105/joss.05153

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90*(3), 414. https://doi.org/10.1037/0022-0663.90.3.414

Hattie, J. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement* (1st ed.). Routledge. https://doi.org/10.4324/9781003380542

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, & Adriane Boyd. (2020). *spaCy: industrial-strength natural language processing in Python*. https://doi.org/10.5281/zenodo.1212303

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

Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J. (2024). Empirische Arbeit: Comparing generative AI and expert feedback to students' writing: Insights from student teachers. *Psychologie in Erziehung Und Unterricht, 71*(2), 80–92. https://doi.org/10.2378/peu2024.art08d

Jensen, L. X., Bearman, M., & Boud, D. (2023). Characteristics of productive feedback encounters in online learning. *Teaching in Higher Education, 30*(1), 69–83. https://doi.org/10.1080/13562517.2023.2213168

Jensen, L. X., Bearman, M., & Boud, D. (2023). Feedback encounters: towards a framework for analysing and understanding feedback processes, *Assessment & Evaluation in Higher Education, 48*(1), 121-134, https://doi.org/10.1080/02602938.2022.2059446

Joyner, D. A. (2017). Scaling expert feedback: Two case studies. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, 71–80. https://doi.org/10.1145/3051457.3051459

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ..., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kettanurak, V. (Nui), Ramamurthy, K., & Haseman, W. D. (2001). User attitude as a mediator of learning performance improvement in an interactive multimedia environment: An empirical investigation of the degree of interactivity and learning styles. *International Journal of Human-Computer Studies, 54*(4), 541–583. https://doi.org/10.1006/ijhc.2001.0457

Lawson, A. P., & Mayer, R. E. (2024). Role of individual differences in executive function for learning From distracting multimedia lessons. *Journal of Educational Computing Research, 62*(3),756-784. https://doi.org/10.1177/07356331231215752

Liu, N.-F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279–290. https://doi.org/10.1080/13562510600680582

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

Malmström, H., Stöhr, C., & Ou, W. (2023). Chatbots and other AI for learning: A survey of use and views among university students in Sweden. *Chalmers Studies in Communication and Learning in Higher Education, 1*(10.17196) https://doi.org/10.17196/CLS.CSCLHE/2023/01

Mayer, R. E. (2002). Multimedia learning. *Psychology of learning and motivation, 41*, 85-139. Academic Press. https://doi.org/10.1016/S0079-7421(02)80005-6

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions.

Computers and Education: Artificial Intelligence, 6, 100199. https://doi.org/10.1016/j.caeai.2023.100199

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

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518

Nielsen, B. L. (2012). Science teachers' meaning-making when involved in a school-based professional development project. *Journal of Science Teacher Education, 23*(6), 621-649. https://doi.org/10.1007/s10972-012-9300-5

Nguyen, N. (2023). Automated feedback coach: Ai Tutor for better feedback writing. https://feedbackfruits.com/blog/automated-feedback-coach-student-ai-tutor-to-write-better-feedback

Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction, 19*(3), 259–271. https://doi.org/10.1016/j.learninstruc.2008.05.002

Nückles, M., Hübner, S., Dümer, S., & Renkl, A. (2010). Expertise reversal effects in writing-to-learn. *Instructional Science, 38*(3), 237–258. https://doi.org/10.1007/s11251-009-9106-9

Ozogul, G., & Sullivan, H. (2009). Student performance and attitudes under formative evaluation by teacher, self and peer evaluators. *Educational Technology Research and Development, 57*, 393-410. https://doi.org/10.1007/s11423-007-9052-7

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational research review, 9*, 129-144. https://doi.org/10.1016/j.edurev.2013.01.002

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

Prilop, C. N., & Weber, K. E. (2023). Digital video-based peer feedback training: The effect of expert feedback on pre-service teachers' peer feedback beliefs and peer feedback quality. *Teaching and Teacher Education, 127*, 104099. https://doi.org/10.1016/j.tate.2023.104099

Prilop, C. N., Weber, K. E., & Kleinknecht, M. (2021). The role of expert feedback in the development of pre-service teachers' professional vision of classroom management in an online blended learning environment. *Teaching and Teacher Education, 99*, 103276. https://doi.org/10.1016/j.tate.2020.103276

Prilop, C. N., Weber, K. E., & Kleinknecht, M. (2019). How digital reflection and feedback environments contribute to pre-service teachers' beliefs during a teaching practicum. *Studies in Educational Evaluation, 62*, 158–170. https://doi.org/10.1016/j.stueduc.2019.06.005

Prins, F. J., Sluijsmans, D. M. A., & Kirschner, P. A. (2006). Feedback for General Practitioners in Training: Quality, Styles, and Preferences. *Advances in Health Sciences Education, 11*(3), 289–303. https://doi.org/10.1007/s10459-005-3250-z

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench—A benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science, 5*(1), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

Ruwe, T., & Mayweg-Paus, E. (2023). "Your argumentation is good", says the AI vs humans – The role of feedback providers and personalised language for feedback effectiveness. *Computers & Education: Artificial Intelligence, 5*, 100189. https://doi.org/10.1016/j.caeai.2023.100189

Ryan, T., Henderson, M., Ryan, K., & Kennedy, G. (2022). Feedback in higher education: aligning academic intent and student sensemaking. *Teaching in Higher Education, 29*(4), 860–875. https://doi.org/10.1080/13562517.2022.2029394

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Strijbos, J.-W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*(4), 291–303. https://doi.org/10.1016/j.learninstruc.2009.08.008

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning & instruction, 4*(4), 295-312. https://doi.org/10.1016/0959-4752(94)90003-5

Topping, K. J. (2021). Digital peer assessment in school teacher education and development: a systematic review. *Research Papers in Education, 38*(3), 472–498. https://doi.org/10.1080/02671522.2021.1961301

Van De Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research, 85*(4), 475-511. https://doi.org/10.3102/0034654314564881

Vygotsky, L. S. (1962). The Development of Scientific Concepts in Childhood. In L. Vygotsky & E. Hanfmann, G. Vakar (Eds.), *Thought and language* (pp. 82–118). MIT Press. https://doi.org/10.1037/11193-006

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child *Psychology and Psychiatry, 17*(2), 89-100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity, 1*, 1–18. https://doi.org/10.1155/2021/8812542

Zottmann, J. M., Stegmann, K., Strijbos, J.-W., Vogel, F., Wecker, C., & Fischer, F. (2013). Computer-supported collaborative learning with digital video cases in teacher education: The impact of teaching experience on knowledge convergence. *Computers in Human Behavior, 29*(5), 2100–2108. https://doi.org/10.1016/j.chb.2013.04.014

Appendix:

TextDescriptives Github link: https://github.com/HLasse/TextDescriptives

vaderSetiment Github link: https://github.com/cjhutto/vaderSentiment

Forfattere

Rasmus R. Hansen

PhD Student

Aarhus University

Centre for Educational Development

Contact: rasmushansen@au.dk

Rikke Frøhlich Hougaard

Educational Developer

Aarhus University

Centre for Educational Development

Karen Louise Møller

Special Consultant

Aarhus University

Centre for Educational Development

Annika Büchert Lindberg

Senior Consultant

Aarhus University

Centre for Educational Development

Tobias Alsted Nielsen

Special Consultant

Aarhus University

Centre for Educational Development

Christopher Neil Prilop

Associate Professor

Aarhus University

Centre for Educational Development

[i] Due to ethical considerations, demographics were calculated based on an anonymous survey and may differ from the actual demographics.

[ii] FeedbackFruits have since further developed and rebranded their AI tools as "Acai", which means that current information and details may not correspond to the AI tool that was used in this study, where data collection ran from December 2023 to June 2024.

[iii] The GitHub links to both TextDescriptives and vaderSentiment can be found in the appendices.