


Teacher, peer, or AI? Comparing effects of feedback sources in higher education

Joshua Weidlich^{a,b,*} , Flurin Gotsch^a, Kai Schudel^a, Claudia Marusic-Würscher^a, Jennifer Mazzarella^a, Hannah Bolten^a, Dario Bütler^a, Simon Luger^a, Bettina Wohlfender^a, Katharina Maag Merki^a

^a Institute of Education, University of Zurich, Freiestrasse 36, 8032 Zurich, Switzerland

^b Educational Research, Zurich University of Teacher Education, Lagerstrasse 2, 8090 Zurich, Switzerland

ARTICLE INFO

Keywords:

Feedback
Artificial intelligence
Peer feedback
Experiment
Large language model
Feedback literacy
Motivational orientation

ABSTRACT

With the emergence of Large Language Models (LLMs), AI-generated feedback is gaining traction as a scalable feedback source for higher education. To qualify as viable alternatives for higher education classrooms, its relative effectiveness compared to teacher feedback and peer feedback needs to be better understood. To this end, a randomized field experiment ($N = 90$) compared the effects of three feedback sources—teacher, peer, and LLM—on students' feedback perceptions and their achievement. To eliminate potential confounds, we (a) controlled for learning gains that may result from students giving feedback to their peers and (b) blinded feedback sources from feedback recipients. Results showed that students rated teacher feedback as less fair and harder to accept than peer and LLM feedback. Students receiving teacher feedback also indicated less willingness to revise their work based on the feedback. Conversely, teacher feedback produced the strongest improvements in scientific argumentation and formal quality of students' work. Here, LLM feedback yielded the smallest improvement overall. Lastly, feedback literacy and intrinsic motivation partly moderated feedback effects on perceptions and achievement outcomes. For example, students with more productive attitudes toward feedback achieved higher argumentation quality after receiving teacher feedback than their peers. Findings indicate that the impact of feedback on both student perceptions and performance depends on the interplay between the feedback source and learner dispositions; deliberately aligning these factors could therefore amplify the benefits of feedback interventions. Future research should explore hybrid and adaptive feedback models that integrate human and AI input.

1. Introduction

Feedback plays a crucial role in fostering student learning and development [1,2]. High-quality feedback can stimulate critical thinking, guide learners to refine their work, and promote self-regulated learning habits [3,4]. Crucially, however, quality feedback—typically provided by instructors—is costly to produce and, therefore, does not scale easily to larger groups of students [5,6]. As a result, higher education instructors often face a tension between offering elaborate but infrequent personalized feedback and providing more frequent but relatively generic feedback [7].

The recent advent of broad-purpose Large Language Models (LLMs) like ChatGPT has altered this landscape by dramatically increasing the

availability of potential feedback. In principle, educators can now generate unlimited, personalized, and near-instantaneous feedback for student work. However, quality assurance remains crucial, given the well-documented inaccuracies and superficialities occasionally produced by LLM-based systems [8,9]. This scalability potential calls for rigorous research comparing the quality of LLM-based feedback with more established feedback sources, such as instructor feedback and peer feedback. Peer feedback, in particular, has long been recognized as a relatively scalable strategy for fostering feedback [10,11]. By scaffolding students to serve as feedback providers for each other, peer feedback systems tap students themselves as potentially rich feedback sources, thereby increasing the volume of formative input available. The relative efficacy of traditional teacher feedback, LLM-generated

* Corresponding author.

E-mail address: joshua.weidlich@ife.uzh.ch (J. Weidlich).

<https://doi.org/10.1016/j.caeo.2025.100300>

Received 23 June 2025; Received in revised form 15 September 2025; Accepted 16 October 2025

Available online 17 October 2025

2666-5573/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

feedback, and peer feedback—both in shaping students' perceptions of feedback quality and in driving actual improvements in learning outcomes—is therefore of critical importance for educational practice. As scalable feedback systems become increasingly viable, understanding the strengths and limitations of each source can help educators design more effective and sustainable feedback ecosystems.

Beyond the strengths and limitations of different feedback sources, students' individual differences are increasingly recognized as critical moderators of feedback effectiveness [12]. Broadly, motivational orientation—whether students are primarily driven by intrinsic interest or extrinsic rewards—can shape how deeply they engage with feedback [13,14]. Intrinsically motivated learners are more likely to process and apply feedback, regardless of its source, while extrinsically motivated students may engage more superficially. In addition, feedback literacy, defined as the skills and dispositions enabling students to understand and act on feedback, represents a feedback-specific competence which influences feedback uptake [15,16]. Considering these factors is essential to understand for whom different feedback modalities are most effective [17,18]. However, empirical evidence about differential effects of feedback sources as a result of individual differences are still missing.

Against this background, this study investigates the effects of three distinct feedback sources—teacher, LLM, and peers—in a higher education lecture course. Extending work from the growing body of research on LLM-based feedback in education (e.g., [8,9,19–23]), we directly compare three feedback sources, offering a broader view of the differential impacts of using these feedback sources. Additionally, we examine not only students' perceptions of feedback—such as fairness, usefulness, and acceptance—but also achievement, which we assessed via objective improvements in the quality of their revised submissions. Finally, we consider the roles of student feedback literacy and motivational orientation as moderators, offering a more comprehensive understanding of how diverse learners engage with different feedback modalities.

2. Literature review

2.1. The role of feedback in learning

Feedback, that is, information about one's learning provided to students, is recognized as a powerful driver of student learning [1,2]. When timely, specific, and personalized, it can help students identify strengths and areas for improvement, narrowing the gap between current and desired performance. Beyond correcting errors, feedback can foster deeper learning by prompting students to critically evaluate their reasoning and refine their work [3].

Particularly effective is formative feedback, which, compared to summative feedback, does not evaluate learning outcomes after the fact but provides learning support while students are working on an assignment or preparing to revise their work. Good formative feedback offers students orientation and guidance, helping them assess their progress and facilitating necessary course corrections [24]. Crucially, however, feedback alone does not guarantee learning; students must productively engage with it, process it deeply, and derive meaningful implications to guide their next steps [16]. This underscores several important factors. First, students' initial perceptions and appraisals significantly influence the depth of their engagement with feedback [25]. Second, feedback can elicit important affective and motivational responses, which shape students' subsequent learning behaviors and thus the overall impacts of the feedback. Evaluations of feedback designs and formats should therefore explicitly consider affective and motivational components to fully determine their educational utility [26,27]. Third, effective feedback is ultimately reflected in improved learning outcomes. Consequently, robust evaluations of feedback effectiveness should incorporate objective measures of student learning rather than relying exclusively on self-report measures [23]. Finally, students differ considerably in how effectively they can utilize feedback. Feedback

literacy—the dispositions and skills enabling students to maximize feedback opportunities—is increasingly recognized as a critical factor influencing feedback effectiveness in higher education [15,16,28].

While these insights predominantly emerged from research on "classic" feedback contexts—typically instructor-provided and not technology-mediated—we suspect they similarly apply to feedback from peers or generated by large language models (LLMs). However, the growing interest in peer-generated and AI-based feedback warrants empirical exploration to examine whether these established principles fully generalize or whether new considerations emerge in these increasingly prevalent educational settings. The following subsections will lay out the state of research into peer feedback and LLM-based feedback, review the recent flurry of comparative feedback studies, and identify key research gaps.

2.2. Peers as a scalable feedback source

Personalized teacher feedback is considered the gold standard of feedback, as instructors' expert knowledge and credibility can yield high-quality guidance directly addressing student needs. However, delivering comprehensive, individualized feedback to each student is time-intensive, and consequently, educators report that workload and time constraints limit their ability to provide timely and frequent feedback [5,6].

In light of these challenges, peer feedback has emerged as a promising complementary approach, which encompasses students assessing and reviewing each other's work. By implementing peer feedback, educators can dramatically increase the volume of feedback information available to each learner. Moreover, a large body of research indicates that peer feedback can be beneficial for recipients and providers [10, 11]. The act of giving feedback encourages students to consolidate their understanding, while receiving peer critiques may be more relatable and accessible, as peers often share a similar perspective and use less formal language [29]. Under these conditions, peer feedback can enhance students' willingness to take up and act on the comments they receive [30]. This combination of scalability and mutual learning has made peer feedback a widely studied and popular strategy for extending feedback opportunities in higher education.

At the same time, the peer feedback process is not without its challenges. The quality of peer critiques can be inconsistent, ranging from detailed and actionable comments to superficial or vague suggestions [31,32]. Peers, by definition, lack subject-matter expertise and evaluative authority, which can lead students to question the credibility of peers' comments, particularly in contexts where only teachers determine final grades [33,34]. In addition, implementing peer feedback requires faculty investment in orienting students, designing rubrics, and monitoring exchanges, which can diminish its appeal in resource-constrained teaching contexts [35].

Existing research highlights several strategies to mitigate these limitations. Structured scaffolds—such as training students in the use of rubrics, providing exemplars, and offering iterative practice—can enhance both the quality and uptake of peer feedback [36]. At the same time, assessing the work of one's peers and giving them feedback is shaped not only by cognitive demands but also by interpersonal dynamics and students' conceptions of the activity [37,38]. In line with this, a review of anonymity effects by Panadero and Alqassab [39] found that anonymous peer assessments can make it easier for students to give critical feedback. Concerns about credibility can be alleviated when peer feedback is positioned as complementary to teacher feedback or integrated into collaborative learning designs where responsibility is shared [40]. Finally, recent comparative work suggests that when structured in this way, peer feedback can rival or even surpass AI-generated feedback in terms of specificity and usefulness for improving academic writing [19].

These findings suggest that while peer feedback has inherent limitations, thoughtful pedagogical design can substantially enhance its

effectiveness. Further, it is a uniquely generalizable and widely applicable feedback approach, due to the possibility of conducting peer feedback as an entirely offline affair. These strengths make it a valuable comparison point when considering the relative strengths of teacher and LLM-based feedback.

2.3. LLM-based feedback as emerging paradigm

AI-based feedback systems have emerged to address the tension between scalability and workload demands of personalized feedback [41], but until a few years ago, many of these systems primarily focused on error detection rather than formative, process-oriented feedback [42, 43]. However, a growing class of more sophisticated systems—such as those developed for collaborative learning in teacher education [27], medical training simulations [44], and academic writing tasks [45]—demonstrate that it is indeed possible to deliver formative, process-oriented feedback at scale. Nevertheless, these systems typically remain domain-specific, posing significant challenges for adaptation across different educational contexts and content areas [46,47].

Given this, LLMs provide significant potential over non-LLM systems, due to their flexibility in processing student inputs as a result of their extensive training on vast datasets. Even general-purpose (“vanilla”) LLMs, which are not explicitly designed with education in mind, in some cases can produce feedback on essay-writing tasks that rivals feedback from teachers [20,23] or peers [19]. Previously, feedback of comparable quality was typically available only through skilled educators or specialized automated systems designed for specific tasks. Now, general LLMs can provide students with immediate, personalized feedback addressing aspects such as scientific argumentation, formal aspects of academic texts, and—depending on the topic and the required depth—also factual correctness of student inputs [20,48]. Crucially, even general LLMs demonstrate a strong grasp of criterion-based quality standards [49].

However, important limitations and concerns remain. A core issue is still whether AI-generated feedback can match the pedagogical quality of human feedback in terms of accuracy, relevance, and specificity [9]. LLMs like ChatGPT may occasionally generate incorrect or overly generic comments because they rely on statistical patterns rather than true content understanding [8]. Moreover, such systems typically lack the empathy and context-awareness intrinsic to human feedback, potentially resulting in comments that feel impersonal or unhelpful [21, 22]. On the other hand, increasingly there seems to be a tendency in LLMs toward sycophancy, that is, systems being overly supportive or even fawning [50,51], which may affect how students like and process feedback. Crucially, current LLMs are not specifically trained to provide pedagogically optimal feedback, and their effectiveness in educational contexts remains uncertain, necessitating careful validation and fine-tuning for classroom use [8,9].

2.4. Existing comparative studies

Given these developments, an increasing number of studies have compared various feedback sources in higher education, focusing primarily on student perceptions and feedback effectiveness. Perceptions are crucial, as even sound and personalized feedback is useless if students fail to process or trust it. Recent studies show that students generally express greater trust and preference for human instructors over AI-generated feedback. For instance, Er et al. [8] demonstrated that students in programming courses valued instructor feedback significantly more than AI feedback, largely due to perceived expertise and motivational support. Li et al. (2024) confirmed this in a flipped classroom setting, showing both higher satisfaction and better performance from instructor feedback compared to GPT-4, again highlighting human motivational and emotional support as critical. Jansen et al. [9] similarly observed significant preferences among student teachers for human expert feedback over AI-generated feedback.

While students often prefer instructor feedback, recent findings on AI-generated feedback indicate it can still offer considerable formative potential, particularly in resource-constrained contexts. Steiss et al. [23] and Dai et al. [20] found that, despite contextual inaccuracies and occasional generality, GPT-4-generated feedback could be highly readable, consistent, and potentially valuable when expert instructor feedback is not readily available. Of note, students’ acceptance of AI feedback can be contingent on idiosyncratic factors. Nazaretsky et al. [22] reported that students initially assessed AI-generated feedback positively but revised their judgements downward after having discovered its source, indicating a source credibility effect. Thus, investigating how students perceive the actual feedback quality would require students to be blinded to feedback sources.

A notable recent study by Usher [52] extends these comparative insights by examining feedback across three feedback sources: instructors, peers, and LLM chatbots. Students reported that chatbots provided detailed, elaborative feedback on their work, offering structured, actionable suggestions occasionally even surpassing instructor- and peer feedback. On the other hand, chatbot-generated feedback included irrelevant or incorrect suggestions at times, requiring critical evaluation by students. Peer feedback, in contrast, was individualized and context-sensitive but often superficial or inconsistent in quality. The instructor feedback demonstrated a high correlation with peer grading, emphasizing human assessors’ capacity for nuanced understanding and context sensitivity, particularly when guided by clear rubrics and some training [52].

Beyond student perceptions or expert ratings, some studies have begun to measure the objective impacts of feedback sources. Meyer et al. [48] showed that LLM-generated feedback improved students’ text revisions and motivation significantly over no feedback. Banihashem et al. [19] found peer feedback particularly effective at identifying specific areas for improvement in essays, contributing to improvements in essay quality. Er et al. [8], however, demonstrated the continued superiority of instructor feedback over AI-generated feedback in improving student performance, emphasizing context sensitivity and personalized guidance as decisive factors.

2.5. Gaps in the current research

Despite the recent increase in comparative studies, several research gaps remain. Most studies assess only two feedback sources (e.g., AI vs. instructor or AI vs. peer), leaving the relative strengths and limitations of teacher, peer, and AI feedback largely unexplored within the same educational context. Although Usher [52] recently conducted a rare three-way comparison, its primary focus was grading consistency and qualitative content analysis rather than objective learning outcomes or moderating student characteristics. Importantly, aside from Nazaretsky et al. [22], existing comparative studies did not blind students to the feedback source, thus possibly confounding students’ genuine perceptions of feedback utility with more idiosyncratic judgements about the source.

This study addresses these gaps by systematically comparing three distinct feedback sources—traditional (teacher), scalable-yet-analogous (peer), and frontier (LLM-based)—to better understand their relative utility for students, while blinding students to the feedback source. Such direct and controlled comparisons can inform educators about how students perceive the quality and utility of such feedback, which may be particularly important for resource-limited educational settings. Given the critical role of student perceptions in shaping feedback engagement in general [25,53,54], our first research question (RQ1) therefore examines these perceptions: *How do students perceive the relative quality of three feedback sources, teacher- vs. LLM- vs. peer-feedback?*

Most comparative studies evaluating feedback effectiveness rely heavily on students’ perceptions, such as perceived usefulness or satisfaction (e.g., [22]). Although valuable for insights into students’ initial reception and appraisal—crucial for understanding their processing and

engagement of feedback [25] self-reports can be susceptible to biases, including social desirability or inaccurate self-assessment [55,56]. We agree with Steiss et al. [23] that objective assessments of feedback impact, that is, uptake of feedback information for improvement, are therefore essential to establish feedback efficacy. While some studies such as Er et al. [8], Li et al. [21], and Meyer et al. [48] include performance outcomes, comparative studies using objective measures to assess learning improvements from teacher, AI, and peer feedback are lacking. Thus, this study further investigates effects of feedback sources on achievement (RQ2): *What is the relative efficacy of three feedback sources, teacher- vs. LLM- vs. peer-feedback, on the revision quality of students' assignments?*

A growing body of research highlights the role of individual differences in shaping how feedback is perceived, processed, and used [6,4,12,57]. With students increasingly viewed as active agents, their ability and willingness to engage with feedback—emotionally, cognitively, and behaviorally—becomes central to its effectiveness [58]. We focus on two complementary moderators linked to feedback uptake: (a) motivational orientation, which broadly influences persistence and depth of engagement [59,60], and (b) feedback literacy, which describes students' ability to interpret and use feedback [15].

Motivational orientation governs how students sustain motivation, either autonomously (intrinsically) or in a controlled, extrinsic way [13,61]. Although often studied as an outcome of feedback [26,48], it can also moderate engagement [62]. Autonomous motivation supports deep engagement, while controlled motivation is associated with surface processing [63]. Thus, intrinsically motivated students are more likely to reflect on and apply feedback, whereas extrinsically motivated students may engage only superficially, profiting less overall and especially from lower-quality feedback. As perceptions of usefulness can mislead learners about what produces durable learning [64], autonomous motivation is particularly important for persisting with demanding feedback. In this context, intrinsically motivated students may be more willing to work with challenging teacher feedback [7], while extrinsically motivated students may gravitate toward more accessible but less deep LLM feedback [8].

Feedback literacy targets the competencies that enable students to understand, value, and use feedback [15,65]. Students vary in this capacity, shaping their emotional and cognitive responses and ultimately how feedback affects learning [16,25]. While widely recognized as central to feedback engagement, its moderating role is still underexplored [66]. Students with higher feedback literacy are more likely to benefit from demanding or critical comments, as they view feedback constructively and can process complex suggestions into revision strategies. In contrast, students with lower feedback literacy may only take up surface-level aspects, leaving them more reliant on accessible feedback—such as that from LLMs, which comparative studies suggest can be more generic [8].

The constructs are complementary: motivational orientation is broad and domain-general, whereas feedback literacy is feedback-specific and task-proximal. Together they allow us to consider both general and more specific predictors of who benefits from different feedback sources. Thus, our last research question (RQ3) is concerned with individual differences of students: *What is the role of motivational orientation and student feedback literacy as potential moderators of these effects?*

Finally, this study deliberately frames feedback as distinct *sources* rather than a pedagogical intervention. One reason for this is that peer feedback typically includes desirable cognitive and metacognitive gains through analyzing peers' work and formulating actionable feedback [10,11]. While a major pedagogical benefit of the peer feedback method, for our purposes, these potential gains obscure the inherent quality and utility of the feedback itself. This source of potential confounding was also noted as a major limitation in the meta-analysis on peer-assessment by Double et al. [67]. By specifically isolating feedback as a source, this study uniquely focuses on the comparative quality, credibility, and efficacy of feedback from teachers, peers, and LLM-generated feedback.

Given the rise of scalable LLM-based systems and ongoing questions about their educational value [20,23], this study can thus offer practical insights for feedback design in resource-limited settings.

3. Method

We created an Open Science Framework (OSF) project for this research study, including a preregistration of our method and analysis, as well as supplemental material, available at: <https://doi.org/10.17605/OSF.IO/EZANP>

3.1. Research design

To address our research questions, this study employed a randomized field experiment with three conditions: Teacher feedback, LLM-feedback, and peer feedback. Thus, the between-subject experimental factor was the feedback source. Field experiments are conducted in authentic learning environments, where the experimental treatment is applied in real-world settings [68], while maintaining—as much as possible—randomized treatment to allow for robust causal conclusions [69]. One key advantage of field studies is their high external validity, meaning the findings are more likely to generalize to actual educational practice than the often artificial conditions of laboratory experiments [70].

As students took part in a whole-class lecture course but also participated in five smaller discussion courses associated with the lecture, we wanted the randomization for this study to ensure that assignment to experimental groups was balanced between these discussion courses. To achieve this, we employed a type of *stratified randomization* (see e.g. [71]), in which randomization was conducted within each discussion course. This design minimizes confounding effects related to group dynamics in the courses while maintaining the focus on feedback effects.

Importantly, we anticipated a potential methodological confound in the peer-feedback condition: students who analyze their peers' work and formulate detailed feedback engage in generative activities with predictable learning benefits, unrelated to feedback reception alone [67,72]. Thus, observed outcomes in peer-feedback groups could partly reflect this cognitive engagement rather than the quality of the feedback received. To address this, we controlled for feedback-giving effects by having *all* students provide feedback on a randomly selected assignment prior to receiving any feedback themselves. Then, for the experimental treatment students in the peer-feedback condition received a subset of random peer-generated comments. This ensured that differences in outcomes could be attributed specifically to the feedback source rather than the act of feedback provision itself.

As another design decision, we implemented blinding. That is, the feedback did not indicate the source of the feedback and thus students were unaware of the feedback source—although, of course, students were informed that they would be receiving feedback from one of the three sources (see section “Procedure”). Blinding ensures that any observed feedback effects were attributable solely to the feedback's content rather than its perceived origin and potential student biases or preconceptions. Nazaretsky et al. [22] findings on how students changed their minds about feedback quality after having learned that it was formulated by AI lends support for this decision.

3.2. Context and sample

The study was conducted at the University of Zurich, within a lecture course called *Bildungsprozesse und Schule* [translates to Educational Processes and School], which included discussion courses, i.e. smaller courses to discuss and deepen the lecture contents. The study focuses on the first of two graded assignments, which requires students to submit a written text demonstrating both their understanding of the course content and their proficiency in scientific argumentation. After receiving

feedback from a feedback source (teacher, peer, or LLM), students had the opportunity to revise and resubmit their text. The topic of the assignment was school quality, school evaluation, and organizational development in schools (see, e.g., [73,74]). Students were tasked with applying the concepts of the past weeks to a fictitious school evaluation, identifying quality factors, delineating development potential, deriving concrete strategies, and reflecting on potential limitations of these strategies. In addition to content-related aspects, the quality of scientific argumentation and compliance with formal requirements including correct APA referencing were explicitly defined as assessment criteria.

The full population of students in this lecture course consisted of 138 bachelor's students, of which 102 filled out the first questionnaire, 98 filled out the second questionnaire, and 90 remained after removing incomplete and double cases. All 90 students also submitted their assignment and a revision after having received the feedback. Thus, after integrating all datasets, the final sample consists of $N = 90$ students. Table 1 provides a summary of demographic information about the student sample. Statistical tests suggested no imbalances between feedback conditions.

We also collected whether students had used AI while working on their assignments. A majority of students (58 %) reported having used systems like ChatGPT in a minor way only (e.g. correcting grammar) while producing their assignment work. More than a third (38 %) reported no usage at all, whereas only very few (6 %) reported more comprehensive usage (e.g. idea creation). No students reported major usage (e.g. producing whole sections via AI).

3.3. Procedure

At the start of the semester, students were informed about the upcoming feedback experiment. A set of slides was developed for both the lecture and discussion courses to outline the study's aims and procedures. Students were told they could opt out of participation, in which case their data would not be stored or analyzed. Participation in the study did not affect the assignment, feedback, or revision procedures, as these were embedded in the regular course structure. Additional study and data privacy details were provided in the first data collection. Students indicated consent by choosing "yes" to proceed with the questionnaire or "no" to end it.

Self-report data were collected at two time points via questionnaires implemented in LimeSurvey. The first (t1) was administered before students received assignment instructions and it covered student demographic information and individual differences variables like motivational orientation and feedback literacy (see Section 3.6). Students were invited to complete it during lecture or discussion sessions, with follow-up reminders via email and the learning platform. Class time in discussion sections was reserved to support completion, and the survey remained open for one week. Afterward, students submitted their

assignments via the learning environment. To generate peer feedback and control for learning effects from feedback writing, each student was required to review and upload feedback on a randomly assigned peer's work. Students were then randomly assigned to receive feedback from a teacher, a peer, or an LLM.

The second questionnaire (t2) was administered after students received their feedback. Here, students first stated whether they had already fully completed the assignment revision prior to the survey (see Section 4.4 for a test of robustness). They then reported on their perceptions of feedback (see Section 3.6) and indicated to which extent they used GenAI for peer feedback and the assignment task. Again, reminders and class time supported participation, and the questionnaire was accessible for ten days. This period overlapped with the one-week revision window during which students updated and re-uploaded their assignments, helping to maximize response rates. Fig. 1 shows the sequence of key study events.

Participation was incentivized by a raffle of 12 gift cards of 20 Swiss Francs each, redeemable at the university mensa and cafeteria. Students were informed that they were entered into the raffle pool by completing both questionnaires, the assignment, and subsequent revision. Students received their gift cards in the last lecture before the end of the semester.

3.4. Feedback instructions and assessment criteria

To ensure consistent feedback quality across conditions, a set of standardized feedback instructions and multiple assessment criteria was developed. All feedback sources received both the instructions and the criteria to guide the feedback process.

The feedback instructions consisted of an initial briefing during the discussion course and a detailed document outlining how to approach the task of writing feedback, including the required features of the feedback messages. Broadly, the instructions were based on established principles of good feedback (e.g., [75,76,1]). Specifically, the instructions stipulated that each quality criterion (see below) should be rated using a three-point response scale and accompanied by a brief written comment for at least three of the nine criteria. These comments were expected to identify areas for improvement and be one to two sentences in length. Additionally, the instructions required a concluding paragraph of 120 to 200 words that synthesized the key strengths and weaknesses of the submission, including any relevant aspects not covered by the individual criteria. To foster a constructive tone, the instructions emphasized the importance of highlighting strengths alongside weaknesses and focusing on suggestions for improvement rather than praise. As further support, a set of example phrases was provided, such as "I liked ...", "I am surprised...", "I wonder...", "I have trouble making sense of...", and "Here, I would expect..."

The assessment criteria structured the content of the feedback by providing a common framework for evaluating the assignments. They

Table 1
Demographic information across feedback conditions.

Variable	Total (N = 90)	Teacher (n = 31)	Peer (n = 30)	LLM (n = 29)	$\chi^2 / F (df)$	p-value
Gender					$\chi^2(4) = 2.19$.70
Women	78 (87 %)	27 (87 %)	26 (87 %)	25 (86 %)		
Men	11 (12 %)	4 (13 %)	3 (10 %)	4 (14 %)		
Other / not stated	1 (1 %)	0	1 (3 %)	0		
Age (years)	23.9	24.2	23.9	23.5	$F(2, 86) = 0.16$.85
Major					$\chi^2(4) = 4.47$.35
Education	34 (38 %)	14 (45 %)	11 (37 %)	9 (31 %)		
Psychology	40 (44 %)	15 (48 %)	12 (40 %)	13 (45 %)		
other	16 (18 %)	2 (6 %)	7 (23 %)	7 (24 %)		
Discussion course					$\chi^2(8) = 4.69$.79
A	15 (17 %)	3 (10 %)	8 (27 %)	4 (14 %)		
B	21 (23 %)	8 (26 %)	6 (20 %)	7 (24 %)		
C	19 (21 %)	6 (19 %)	7 (23 %)	6 (21 %)		
D	20 (22 %)	8 (26 %)	6 (20 %)	6 (21 %)		
E	15 (17 %)	6 (19 %)	3 (10 %)	6 (21 %)		

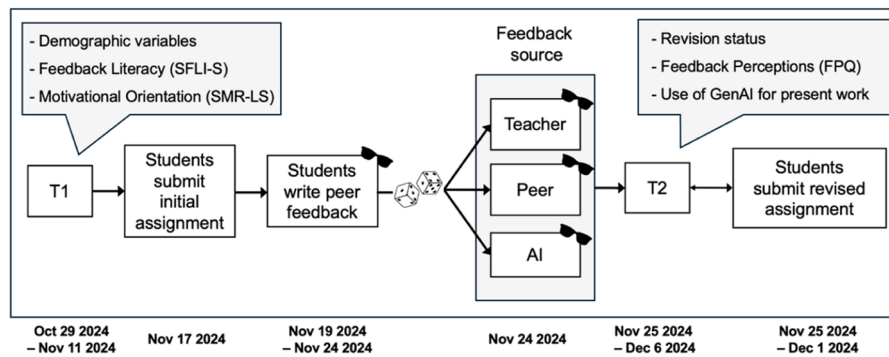


Fig. 1. Sequence of key events for this study. Dice indicate randomization of students to feedback sources; sunglasses indicate blinding of peer feedback and feedback sources.

included aspects such as appropriate and accurate application of course content, coherence of arguments, adherence to scientific writing standards, correct use of APA style, and formal aspects like length and formatting—nine criteria in total. Each criterion was rated on a three-point scale: fully satisfied (2), somewhat satisfied (1), or not at all satisfied (0). Each criterion also allowed for an optional free-text comment. A designated section at the bottom of the document was reserved for the concluding paragraph.

3.5. Feedback sources

3.5.1. LLM feedback

Students in this condition received feedback generated by a custom GPT-4 system (OpenAI, 2024). The model was provided with a curated course knowledge base (lecture slides, readings, assignment brief, and rubric, as well as summaries of lecture transcripts reviewed by the instructor) to ensure relevance and accuracy. For each student submission, the system was prompted to deliver constructive feedback aligned with the assignment criteria and the formal requirement for structure and tone of the feedback. After students had submitted their assignments via the learning platform, a researcher anonymized the submissions before inputting them into the custom GPT. For each new submission, a chat window was opened to avoid cross-contamination between submissions. The model generated feedback based on the provided instructions and assessment criteria. To ensure a consistent format and appearance, the output was manually transferred into the standardized feedback template. The approach to creating LLM feedback, including the complete prompt, knowledge base, and technical setup, are reported in Appendix A.

3.5.2. Peer feedback

As noted in “Research Design”, all students wrote feedback on a randomly assigned peer’s submission, allowing control for potential feedback-giving effects across conditions. Only students in the peer feedback condition, however, received a peer-generated feedback message. The feedback process was anonymized in both directions: students did not know whose work they were reviewing, nor the identity of the peer who reviewed theirs. Peer feedback was assigned by randomly linking students into unidirectional pairs, such that each student appeared once as a feedback giver and once as a potential feedback recipient. These pairings were distributed across discussion courses to eliminate existing social dynamics and reducing the likelihood of students recognizing each other’s work. Students had one week to submit their peer feedback via the learning platform. A research assistant then used the master linking list to allocate peer feedback to the appropriate recipients in the peer feedback treatment group. Students were not instructed to use AI support when writing peer feedback. As they worked on this task from home, however, they might have used any digital tools as they might in their normal studying. At the end of the activity, they

reported the extent to which they had used GenAI for producing peer feedback. Most of students (78 %) reported no usage of Generative AI while formulating peer feedback. A minority (21 %) reported minor usage (e.g. grammar improvement), and only one person reported more comprehensive usage (generation of ideas).

3.5.3. Teacher feedback

Instructors provided feedback using the same criteria and instructions as the other feedback sources. While the basic process reflected standard practice from previous semesters, the introduction of standardized criteria and guidance brought more structure to the task, as instructors did not adhere to detailed standardized feedback instructions or assessment criteria in previous iterations of the course. Teachers had one week to complete their feedback, with roughly 30 min to dedicate to one submission.

All students, regardless of condition, received their written feedback seven days after submitting their assignments.

3.6. Measurement instruments

Key variables were measured via self-report instruments validated for German language contexts (see Table 2).

Feedback literacy was measured with the two-dimension Student Feedback Literacy Instrument–Short (SFLI-S; [66,78]), assessing feedback attitudes and practices. The short form was selected for survey efficiency and because it is currently the most parsimonious scale available. Importantly, like Dawson et al. [79], it captures student behaviors as well as attitudes, thereby covering feedback literacy beyond knowledge or attitudes alone

For motivational orientation, this study measures only the two endpoints of the self-determination continuum to first establish potential moderator effects using the clearest possible contrast in motivational orientations. This was done using the *Skalen zur motivationalen Regulation beim Lernen im Studium* (SMR-LS) due to its desirable psychometric properties in German language [77].

For feedback perceptions, we used the FPQ instrument by Strijbos et al. [34] and Strijbos et al. [80], as it captures various relevant dimensions of how students may perceive feedback, has undergone ample validation, and provides a German translation.

Overall, instruments showed sufficient internal consistency, with the exception of feedback attitudes, which yielded a Cronbach’s alpha slightly below 0.7. This may partly stem from the fact that the SFLI-S is a short-form version of a longer instrument [78], where the four items cover heterogeneous content (i.e. the facets *readiness*, *agency*, *appraisal*, *model*), as compared to more homogeneous dimensions like intrinsic motivation [81].

Table 2
Measurement instruments.

Construct	Dimensions (# of items)	Example	T1 Mean (SD)	T2 Mean (SD)
Feedback Literacy SFLI-S, [66]; 7-point Likert	Attitudes (4)	“I believe that I can contribute to the value of feedback processes.”	$\alpha = 0.68$ M = 5.64 (0.66)	–
	Practices (7)	“When I receive feedback, I carefully take note of every comment.”	$\alpha = 0.81$ M = 5.23 (0.8)	–
Motivational Orientation SMR-LS, [77]; 7-point Likert	Intrinsic (3)	“I’m really enjoying my studies.” (transl.)	$\alpha = 0.89$ M = 4.93 (1.24)	–
	Extrinsic (3)	“I study primarily because I want to earn an academic degree.” (transl.)	$\alpha = 0.74$ M = 4.42 (1.37)	–
Feedback perceptions FPQ, [34]; 10-point Likert	Fairness (3)	“I would consider this feedback justified.”	–	$\alpha = 0.93$ M = 7.44 (2.2)
	Usefulness (3)	“I would consider this feedback useful.”	–	$\alpha = 0.94$ M = 6.63 (2.38)
	Acceptance (3)	“I would accept this feedback.”	–	$\alpha = 0.83$ M = 8.1 (1.89)
	Willingness (3)	“I would be willing to improve my performance.”	–	$\alpha = 0.73$ M = 7.78 (1.53)

3.7. Coding of the quality of students’ work

A total of 90 initial and revised assignments were coded for quality by four coders. Codes were derived from the set of assessment criteria which underlies the feedback provision (see Section 3.4). This set encompassed 9 criteria, including the factual correctness of content aspects, soundness of argumentation, academic writing style, and APA-consistent citations and referencing. Our aim was to evaluate the quality of students’ work with the same criteria that structured the feedback. With this alignment, we can plausibly attribute any improvements in the revisions to the quality and content of the feedback students received, rather than to additional criteria not specified in the task or feedback.

While established coding schemes for scientific argumentation (e.g. [82,83]) offer fine-grained structural analysis, fully implementing such schemes would have required substantial coder training and time, and, critically, would have weakened correspondence with course learning outcomes and rubric guiding the feedback. Our rubric-aligned dimensions—content, argumentation, and formal quality—remain conceptually aligned with this literature: Argumentation emphasizes coherence between claims and evidence and the explicitness of warrants and qualifiers; content refers to disciplinary coverage and accuracy; formal quality captures genre conventions like APA compliance and formal rigor.

For each criterion, we defined three quality levels: inadequate or not present (0), adequate (1), high quality (2). To assign these codes reliably, multiple iterations of discussion and practice coding were required, through which an operationalization for each criterion was developed. For example, distinctions between adequate and high-quality codes were discussed and encoded with examples and if-then rules. For the criteria pertaining to factual correctness of content, these rules needed to be partly content-specific, which requires multiple iterations of practice and discussion. Other discussion topics included

how to code cascading errors, operationalizations of stringency in argumentation, thresholds for typos and grammatical errors, and others. In the end, the criteria were merged into three overarching quality dimensions: Content quality (e.g., correct mapping of development strategies to identified quality deficits), argumentation quality (e.g., stringent argumentation), and formal quality (e.g., APA style). See Table 3 for an overview of quality dimensions.

Four raters coded a subset of twelve assignments to assess interrater agreement on quality features using Cohen’s kappa. Agreement was satisfactory for argumentation ($\kappa = 0.65$) and formal quality ($\kappa = 0.77$), but insufficient for content quality ($\kappa = 0.53$). Closer inspection revealed that one specific content criterion accounted for most of the disagreement. Although this criterion depended on the correctness of a preceding one, the coding team had agreed not to penalize cascading errors—students were to receive credit for a correct response, even if it built on a previous mistake. Despite this guideline, discrepancies indicated that raters interpreted it inconsistently, differing in their tendency to award full points when earlier criteria were not fully met. Given this inconsistency and the lack of resources for further rater training, the problematic criterion was omitted. Following its removal, interrater reliability for content quality increased to $\kappa = 0.71$.

3.8. Analytical approach

We addressed our questions with three model sets. The feedback perception models (RQ1) test whether students’ quality ratings differ by feedback source. The revision quality models (RQ2) predict revised-submission quality from feedback source, adjusting for initial scores. The individual difference models (RQ3) add individual-difference predictors and their interactions with feedback source to evaluate whether they improve fit and explain variation in effects. By fitting these in sequence, rather than forcing a single omnibus model, we first establish whether scalable feedback sources differ from the teacher “gold standard” on average. Only once average effects are clear do we probe for whom those effects vary, yielding both general and more nuanced insights.

For the perception models (RQ1), significant Levene’s and Breusch-Pagan tests indicated violations of the homogeneity of variances assumption. To account for this, we conducted robust ANOVAs using a modified one-step M-estimator with Mahalanobis distances and 1000 bootstrap samples. For each feedback quality perception, we estimated a model with feedback source as categorical predictor and the respective perception as dependent variable. Bootstrapping was used to obtain empirical p-values and to construct confidence intervals for group comparisons, which supports robust inference despite heteroscedasticity and potential deviations from normality. Post-hoc tests yield the unstandardized ψ (psi-hat) effect size metric, but, using winsorized means ($M_{wins.}$) and standard deviations, this can be translated into a standardized metric like Cohen’s d , with thresholds of 0.3 (small), 0.5 (medium), and 0.8 (large; [84]) to make effects comparable between all models. As we are not aware of a power analysis procedure for models with the above specifications, we conducted an ANOVA-based (fixed effects, one-way) sensitivity power analysis ($\alpha = 0.05$, $\beta = 0.80$) as approximation. For the available sample size ($N = 90$) the study was powered to detect effects of $f = 0.33$ for the analyses of feedback quality perceptions (medium to large effects)

For the achievement models (RQ2), homogeneity of variances, normality of residuals, and Q-Q plots suggested minor deviations for argumentation quality, but no substantial violations overall. Thus, we proceeded with the General Linear Models (GLMs) for these analyses. Models for each quality score included feedback source (simple-coded) as predictor, quality dimension (after revision) as dependent variable,

Table 3
Assignment quality dimensions and initial versus revised scores.

Dimension (# of subcriteria)	Content (max. possible points)	Cohen's kappa	Initial Mean (SD)	Revision Mean (SD)	Paired-samples T-test
Content (3)	The coverage and correctness across content aspects (18)	$\kappa = 0.71$	$M = 12.04$ (3.52)	$M = 13.17$ (3.37)	$t(89) = 5.81$ $p < 0.001, d = 0.61$
Argumentation (2)	The overall soundness of the argumentation and use of academic writing style (4)	$\kappa = 0.65$	$M = 2.56$ (1.16)	$M = 2.89$ (1.05)	$t(89) = 4.98$ $p < 0.001, d = 0.52$
Formal (3)	Correctness of APA Style, errors and typos, and formal aspects (e.g. word limit) (6)	$\kappa = 0.77$	$M = 4.16$ (1.30)	$M = 4.49$ (1.27)	$t(89) = 3.74$ $p < 0.001, d = 0.39$

and the initial quality score as centered covariate to control for pre-feedback differences. Significant Omnibus tests (ANOVA) were followed by post-hoc Tukey-adjusted comparisons. Effect sizes are reported as Cohen's *d*. Power analysis (repeated measures ANOVA, between factors) with a $r = 0.8$ correlation between repeated measures yielded a detectable effect size of $f = 0.32$ for the quality analyses (medium to large effects).

The individual differences models (RQ3) encompass a two-step procedure. To gauge whether the additional student variables added to the model's R^2 , we first conducted model comparisons. The nested model mirrored the previous analyses, whereas the full model added moderator variables and their interactions with feedback source. For feedback literacy, the moderators were feedback attitudes and practices; for motivational orientation, intrinsic and extrinsic motivation. Thus, the full model included four additional predictors. A significant increase ΔR^2 prompted further analyses. This was powered (R^2 increase in linear multiple regression) to detect $f^2 = 0.14$, a medium-sized effect. Significant interactions were probed via simple slopes at Mean ± 1 SD. Effect sizes for these analyses are reported as partial η^2 , with thresholds of 0.01 (small), 0.06 (medium), and 0.14 (large; [84]). The models (ANOVA, fixed effects, special, main, and interactions) were powered to detect effect sizes of $f = 0.32$ (medium to large).

4. Results

4.1. Feedback perceptions

Student perceptions of feedback quality as a function of feedback source are shown in Fig. 2.

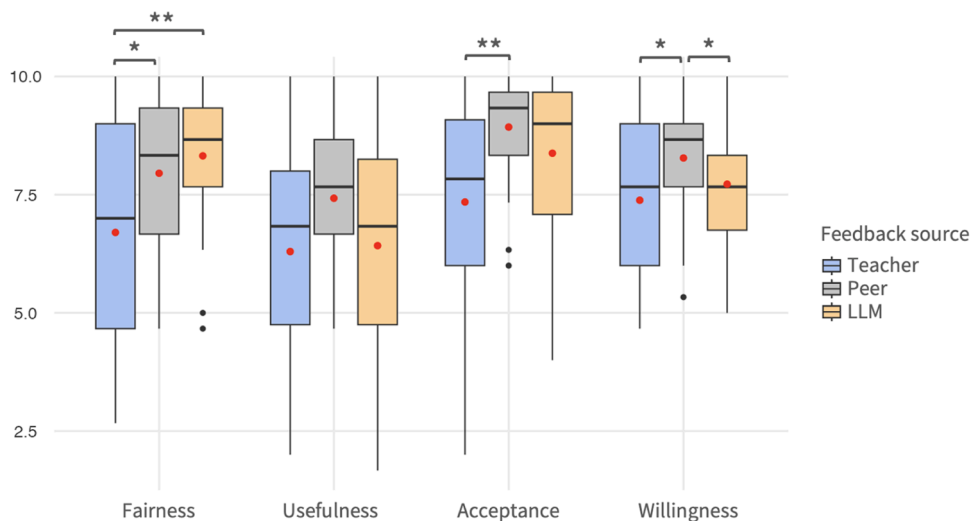


Fig 2. Box plots of student perceptions of feedback quality across four dimensions as a function of feedback condition. Red dots represent mean values, and horizontal lines represent the median. * $p < .05$; ** $p < .01$.

For fairness perceptions, the feedback factor was significant, $F = 3.73, p = .033$. Post-hoc tests showed a significant advantage of LLM feedback ($M_{wins.} = 8.49, SE = 0.32$) over teacher feedback ($M_{wins.} = 6.69, SE = 0.61$), $p = .004, \psi = 1.79$, translated to Cohen's $d = 0.68$. The peer feedback ($M_{wins.} = 8.07, SE = 0.34$) was also rated as significantly fairer than the teacher feedback, $p = .034, \psi = 1.49$, translated to Cohen's $d = 0.56$. The model for usefulness perceptions did not yield a significant feedback factor, $F = 1.52, p = .25$. For acceptance, the feedback factor was significant, $F = 5.16, p = .021$. Post-hoc tests suggested a significant advantage of peer feedback ($M_{wins.} = 9.08, SE = 0.16$) over teacher feedback ($M_{wins.} = 7.6, SE = 0.46$), $p = .002, \psi = 1.47$, translated to Cohen's $d = 0.81$. The remaining comparisons were not significant. Lastly, there was a significant effect of feedback source for willingness perceptions, $F = 3.8, p = .046$. Post-hoc tests suggested a significantly higher willingness for students having received peer feedback ($M_{wins.} = 8.39, SE = 0.23$) than those with teacher feedback ($M_{wins.} = 7.37, SE = 0.48$), $p = .04, \psi = 1.03$, translated to Cohen's $d = 0.53$. Peer feedback also led to higher willingness compared to LLM-feedback ($M_{wins.} = 7.55, SE = 0.26$), $p = .044, \psi = 0.84$, translated to Cohen's $d = 0.64$.

4.2. Effects on achievement

The following analyses assess achievement effects via the revision quality along three dimensions (content, scientific argumentation, formal) as a function of feedback condition, while controlling for the quality of the initial submission.

For content quality, the omnibus tests yielded no main effect of feedback source, $F(2, 83) = 0.53, p = .591$, a significant covariate, $F(1, 83) = 169.08, p < .001, \eta^2 p = 0.69$, and no significant interaction, F

(2,83) = 0.14, $p = .87$.

For argumentation quality, the omnibus test yielded a main effect of feedback source, $F(2, 84) = 4.02, p = .009, \eta^2p = 0.11$, a significant covariate, $F(1, 84) = 205.14, p < .001, \eta^2p = 0.71$, and no significant interaction, $F(2, 84) = 1.71, p = .14$. Tukey-adjusted comparisons showed a significant difference between teacher feedback ($M = 3.04, 95\% \text{ CI } [2.84, 3.24]$) and LLM feedback ($M = 2.59, 95\% \text{ CI } [2.38, 2.79]$), $t(84) = 3.15, p = .006$, Cohen's $d = 0.82$. The remaining comparisons were not statistically significant.

For formal quality, the omnibus test yielded a main effect of feedback source, $F(2, 84) = 3.91, p = .024, \eta^2p = 0.09$, a significant covariate, $F(1, 84) = 170.19, p < .001, \eta^2p = 0.67$, and no significant interaction, $F(2, 83) = 0.11, p = .90$. Tukey-adjusted comparisons showed a significant difference between peer feedback ($M = 4.65, 95\% \text{ CI } [4.39, 4.92]$) and LLM feedback ($M = 4.17, 95\% \text{ CI } [3.89, 4.45]$), $t(84) = 2.47, p = .041$, Cohen's $d = 0.65$. Further, teacher feedback ($M = 4.64, 95\% \text{ CI } [4.37, 4.91]$) led to higher formal quality than LLM feedback, $t(84) = 2.39, p = .049$, Cohen's $d = 0.63$. See Fig. 3 for plots of three quality scores as a function of feedback source.

4.3. The role of individual differences

4.3.1. Feedback literacy

4.3.1.1. Moderation of feedback quality perceptions. To assess the role of feedback literacy as a potential moderator of feedback quality perceptions, we conducted model comparisons upon introducing feedback literacy dimensions. Thus, the nested model included the feedback source, and the full model consisted of the feedback source, both feedback literacy dimension, as well as their interaction with feedback source.

For fairness perceptions, feedback literacy significantly enhanced model fit ($\Delta R^2 = 0.14, p = .027$), due to significant interactions with feedback attitudes, $F(2,78) = 5.24, p = .007, \eta^2p = 0.11$, and feedback practices, $F(2,78) = 3.82, p = .026, \eta^2p = 0.09$, both medium effects according to Cohen [84]. Further probing the interaction revealed that students with -1SD and average feedback attitudes rated the teacher feedback significantly less fair than peer- or LLM-feedback, $F(2,80) = 12.01, p < .001, \eta^2p = 0.23$ (large effect). This was still true for students with average feedback attitudes, $F(2,80) = 8.43, p < .001, \eta^2p = 0.17$ (large effect), but the effect disappeared with +1SD feedback attitudes, where there was no difference between feedback sources (see Fig. 4, left). For feedback practices, a similar pattern of results emerged (see Fig 4, right). Students rated the teacher feedback as significantly less fair at moderator levels of -1SD, $F(2,81) = 10.77, p < .001, \eta^2p = 0.21$ (large), and average levels, $F(2,81) = 7.98, p < .001, \eta^2p = 0.16$ (large). When

students had above average productive feedback practices, no differences in fairness perceptions were observed. For the remaining feedback perceptions—usefulness, acceptance, willingness—feedback literacy did not contribute to better model fit.

4.3.1.2. Moderation of achievement effects. To investigate the role of feedback literacy as a potential moderator of the effects of feedback on revision quality, we conducted model comparisons upon introducing feedback literacy dimensions. Thus, the nested model included the feedback source, and the full model consisted of the feedback source, both feedback literacy dimensions, as well as interaction terms between feedback source and feedback literacy dimension.

For the argumentation quality of students' work, feedback literacy significantly enhanced model fit, $\Delta R^2 = 0.05, p = .009$. The interaction with feedback attitudes was significant, $F(2,80) = 7.95, p < .001, \eta^2p = 0.17$, a large effect [84]. Probing of the interaction revealed that with average and +1SD feedback attitudes benefitted significantly more from teacher feedback in enhancing argumentation quality of their work. This effect was particularly pronounced for students with +1SD feedback attitudes, $F(2,80) = 12.31, p < .001, \eta^2p = 0.24$ (see Fig 5, left), and decreased with lower values of the moderator, then disappearing for -1SD feedback attitudes. Descriptively, a converse pattern of effects appears for feedback practices, but this was not statistically significant $F(2,80) = 1.97, p = .15$ (see Fig 5, right). Here, the effect of teacher feedback depends on feedback practices, but students with -1SD of the moderator benefitted more from teacher feedback than their fellow students with more productive feedback practices. For the content quality and formal quality of student's work, feedback literacy did not enhance model fit.

4.3.2. Motivational orientation

4.3.2.1. Moderation of feedback quality perceptions. To assess the role of motivational orientation as a potential moderator of feedback quality perceptions, we conducted model comparisons upon introducing motivation as predictor. Thus, the nested model included the feedback source, and the full model consisted of the feedback source, extrinsic and intrinsic motivation variables, as well as their respective interaction with feedback source.

The analysis yielded a significantly increased model fit for willingness perceptions, $\Delta R^2 = 0.19, p = .009$. Specifically, there was a significant interaction of intrinsic motivation with feedback source, $F(2,72) = 7.75, p < .001, \eta^2p = 0.18$. Further probing yields a pattern of results where it is only the LLM-feedback for which students' willingness to revise their work was moderated by intrinsic motivation (see Fig 6, left).

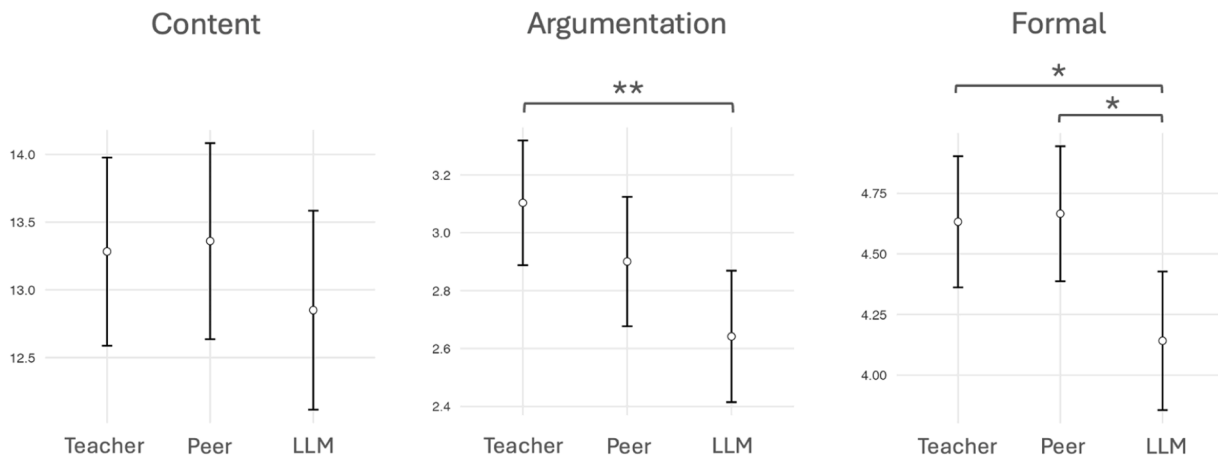


Fig 3. Estimated marginal means of quality dimension by feedback source. Estimates are controlled for initial submission quality. Error bars are 95 % confidence intervals.

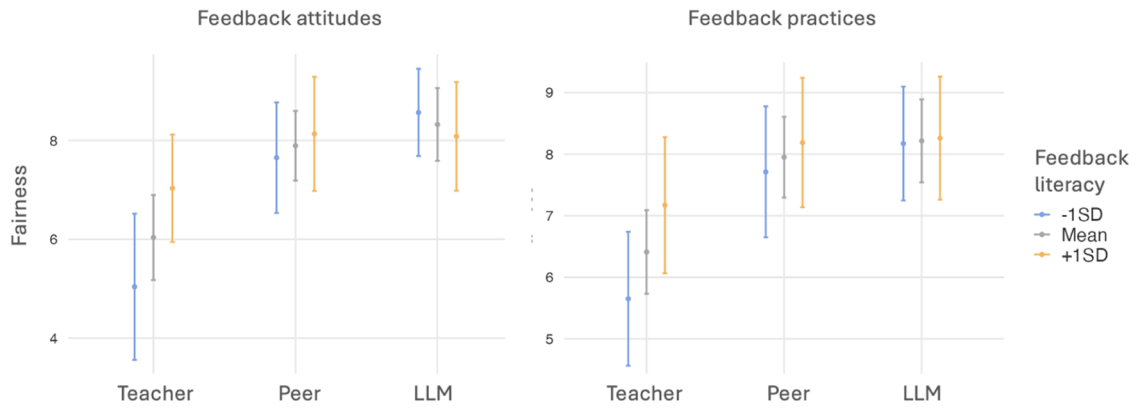


Fig 4. Visualization based on estimated model means for fairness as a function of the interaction between feedback source and feedback attitudes (left) and feedback practices (right).

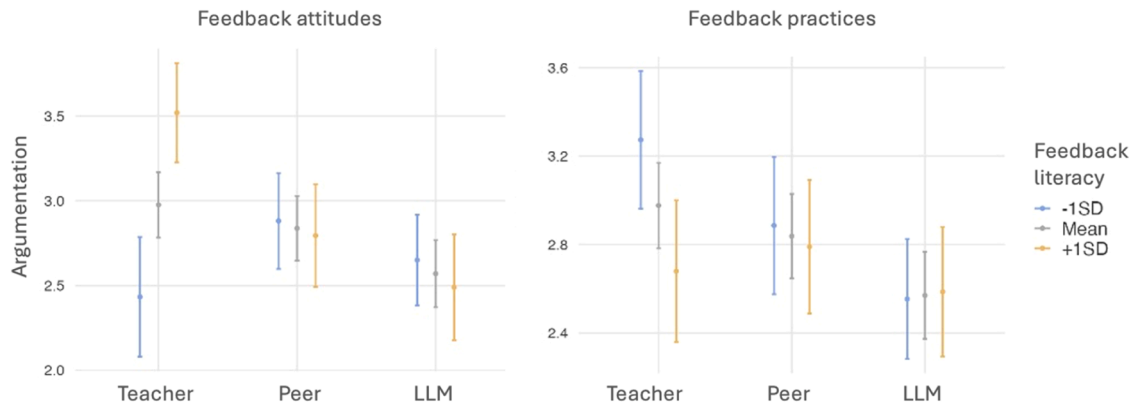


Fig 5. Visualization of estimated model means for scientific argumentation quality score as a function of the interaction between feedback source and feedback attitudes (left) and feedback practices (right).

Simple effects revealed that students with +1SD intrinsic motivation were more willing to revise their work after receiving LLM-feedback, $F(2,72) = 6.12, p = 0.004, \eta^2 p = 0.15$, whereas the opposite was true for students low (-1SD) on intrinsic motivation, $F(2,72) = 4.15, p = 0.02, \eta^2 p = 0.10$. Students with average degrees of intrinsic motivation did not differ between feedback sources. Extrinsic motivation did not yield any such patterns, see Fig 6 (right) for comparison. For the remaining feedback quality perceptions, motivational orientation did not play a moderating role.

4.3.2.2. *Moderation of achievement effects.* Using the same approach as reported above, no significantly increased model fit was observed for all dimensions of the revision quality. Thus, motivational orientation did not influence these effects of feedback sources.

4.4. Tests for robustness

To evaluate whether our results are robust to arbitrary analytical decisions, we report two robustness checks. The results of both sets of

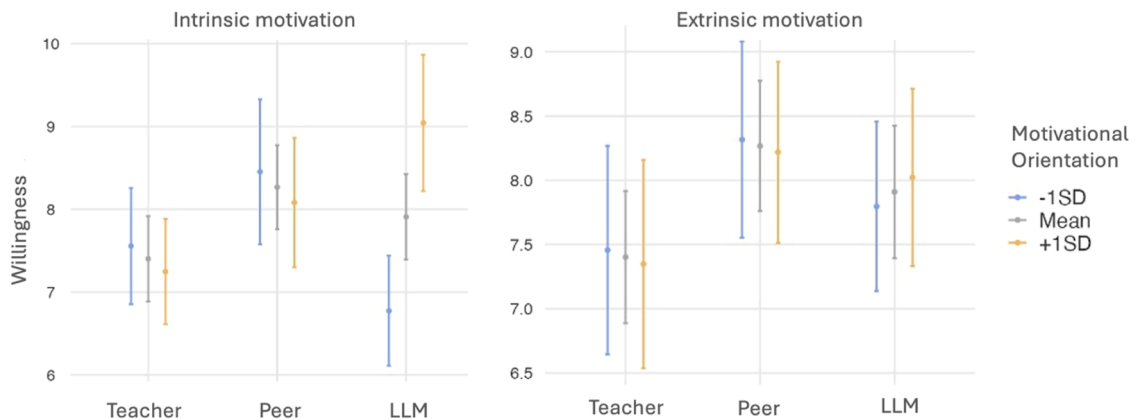


Fig 6. Visualization based on estimated model means for willingness perceptions as a function of the interaction between feedback source and intrinsic motivation (left) and extrinsic motivation (right).

analyses can be found in the supplementary material at <https://doi.org/10.17605/OSF.IO/EZANP>.

First, students participated in discussion courses throughout the semester, resulting in clustered data where students within the same course may share similar learning experiences. Although multilevel models typically address this structure, our sample size was insufficient for such analyses, and we had no substantive interest in the clustering itself. Therefore, we did not include clustering in our primary analyses but conducted robustness checks by adding course membership as a predictor. Results comparing models with and without this predictor showed no meaningful differences in outcomes or model fit.

Due to practical constraints, we could not control whether students completed the second questionnaire before, during, or after revising their work. Most students (85 %) had finished their revision, while 14 % had started but not completed it; only one student had not begun. Plausibly, the utility of feedback becomes particularly salient when students engage with it during revision, as they are more likely to implement feedback they perceive as understandable and agreeable [85]. Thus, as an additional robustness check, we re-analyzed our models including only students who had fully completed their revisions ($n = 72$). These comparisons revealed consistent directional findings despite minor changes in statistical significance due to reduced power.

5. Discussion

This study compared the effects of teacher, peer, and LLM-based feedback on students' perceptions of feedback quality and on improvements upon revising their work after feedback. In addition, we explored the moderating roles of feedback literacy and motivational orientation. Our findings extend prior research by offering a direct, controlled three-way comparison across feedback sources in an authentic higher education setting.

5.1. Summary of findings

Students' perceptions of feedback quality varied significantly across feedback sources. Peer and LLM feedback were perceived as fairer and, in the case of peer feedback, more acceptable than teacher feedback. However, no significant differences emerged regarding perceived usefulness or willingness to use the feedback, except when considering moderating effects. Feedback literacy moderated several perception outcomes: students with higher feedback literacy rated teacher feedback as fairer and more acceptable, suggesting that feedback literacy fosters a more favorable reception of expert feedback. Moreover, intrinsic motivation moderated willingness to use feedback, particularly boosting willingness in the LLM feedback condition.

Regarding objective outcomes, students in all feedback conditions showed improvements in their revised submissions. However, some significant differences emerged: teacher feedback led to stronger improvements in scientific argumentation quality compared to LLM feedback, while peer feedback outperformed LLM feedback on formal writing aspects. Content quality improvements did not differ significantly across feedback sources. Importantly, moderation analyses revealed that students with higher feedback literacy particularly benefited from teacher feedback in terms of argumentation quality, while no such moderation was observed for peer or LLM feedback. Motivational orientation did not moderate learning outcomes.

5.2. Interpretation of results

The finding that peer and LLM feedback were perceived as fairer and more acceptable than teacher feedback challenges traditional assumptions about the superiority of expert feedback, at least from a student perspective. One plausible interpretation is that peer and LLM feedback, possibly due to their language tone or focus, may be perceived as less judgmental and more relatable or encouraging compared to more

authoritative teacher comments. This aligns with prior research emphasizing the relational closeness and linguistic accessibility of peer feedback [29,30] and observations that LLM feedback can appear neutral and objective [20,23], but also supportive, to the point of sympathy [50,86].

The absence of a significant difference in perceived usefulness across feedback sources contrasts with earlier findings emphasizing greater student satisfaction with human feedback [8,21]. This discrepancy might stem from our study's blinding of feedback sources, which likely minimized source credibility biases [22]. While blinding is likely a key reason for convergence of perceived usefulness across sources, other potential mechanisms may also have contributed. First, the design of all three feedback sources was deliberately aligned: each used the same rubric, template, and neutral academic tone, and the LLM was grounded in course materials. Further, students may have a limited ability to differentiate feedback quality without extended engagement. When all feedback is structured and blinded, students may weight surface accessibility (e.g. clarity, concreteness) more to judge usefulness. Thus, perceptions may converge because all feedback "looked" usable, even though its capacity to improve scientific argumentation differed.

Achievement effects, as measured by the revision quality of students' work, tell a more nuanced story. Teacher feedback produced the greatest gains in scientific argumentation quality, confirming previous findings that human feedback better addresses higher-order skills requiring nuanced judgment [8,52]. LLM feedback lagged behind both teacher and peer feedback on the quality dimension of scientific argumentation, reinforcing concerns about its occasional superficiality and lack of deep context understanding [8,9]. Notably, peer feedback was as effective as teacher feedback in improving formal aspects of writing, highlighting the potential of peers to support surface-level textual improvements, a finding consistent with earlier studies [19,10]. This pattern underscores that feedback sources vary in their comparative strengths.

The moderating role of feedback literacy provides additional insights. Students with higher feedback literacy rated teacher feedback more favorably and benefited more from it in improving their argumentation quality. This finding echoes theoretical models of feedback uptake [16] and further supports that feedback literacy is critical for extracting maximum value from high-quality but demanding feedback [15,65]. Conversely, students with lower feedback literacy may find peer or LLM feedback more accessible and easier to engage with, though potentially at the cost of deeper learning gains.

In addition, motivational orientation—particularly intrinsic motivation—moderated willingness to act on feedback, but only in the LLM condition. Students high in intrinsic motivation showed higher willingness to revise based on LLM feedback, whereas students low in intrinsic motivation showed less willingness. This interaction indicates that intrinsically motivated students may be better equipped to productively engage with the affordances and limitations of LLM-generated feedback [62], a finding of practical importance given the increasing role of AI tools in self-directed learning.

5.3. Implications for practice

5.3.1. Addressing common challenges

Our findings, together with our design choices, speak to several frequently discussed challenges. For peer feedback, variability and credibility concerns are well-documented (see e.g., [10,33]). We attempted to stabilize quality and tone through standardized criteria, a shared template and exemplary phrasings, and we used blinding to restrict interpersonal effects [37,38]. To circumvent the confound—indeed a benefit for the classroom—that giving feedback can improve learning [67], all students authored feedback before receiving any. This ensures that our outcome differences reflect the source of feedback rather than the act of writing feedback. Despite the absence of review training, students were able to produce helpful feedback on their peer's work [10,30]. In light of the feedback source blinding, the

positive perceptions toward peer feedback point to reliability and accessibility of the messages [29].

For LLM feedback, concerns include inaccuracies, superficiality, and limited pedagogical nuance [20,8,9]. To mitigate lack of context, we grounded the LLM in course materials and assessment criteria while constraining output format. This likely contributed to positive student reactions to it, yet it was outperformed by both peer feedback and teacher feedback in terms of revision quality, most strikingly by teacher feedback on argumentation quality. This aligns with comparative findings of human expertise yielding deeper disciplinary insights and tailored guidance [8,21].

This echoes prior theoretical arguments that the choice of feedback sources should align with the nature of the learning objectives [87,24]. Where deep disciplinary reasoning and conceptual change are targeted, investing teacher's time into personalized feedback may yet remain indispensable [6]. Conversely, when the learning goals are related to technical writing, structure, or lower-order skill consolidation, scalable feedback mechanisms like peer- or AI-generated feedback can be deployed effectively to provide timely and voluminous support without overburdening educators [5,10]. Thus, a differentiated feedback architecture that strategically combines multiple feedback sources depending on task demands could offer a promising way forward to mitigate challenges of single feedback sources, which is particularly important for resource-constrained educational environments. Approaches to such differentiated architectures can draw on hybrid human-AI design principles that allocate roles between teachers, students, and AI systems [88]. In writing-intensive contexts, concrete hybrid models like PAIRR braid peer review with AI-generated critique and structured reflection, illustrating how to distribute feedback work while preserving pedagogical intent [89].

Notably, students' perceptions did not fully align with objective learning outcomes in our study: Peer and LLM feedback were judged as fair and subjectively easy to accept, yet teacher feedback produced the strongest gains in scientific argumentation. This divergence underscores that perceived utility does not necessarily map onto improvement in performance ([90,64,91]; see also recently [92]). At the same time, in feedback contexts some self-reports are not mere proxies but *ground truth*: appraisals such as perceived fairness, tone, and acceptance directly shape whether students engage with, process, and implement comments. Thus, even pedagogically strong feedback can have little effect if students resist it for subjective reasons; in such cases the subjective appraisals become proximal determinants of uptake and hence impact (see e.g. [58]). In the present study, we suspected source-credibility biases to be an important influence on feedback perceptions (cf [22]). Thus, we blinded feedback sources, which may help explain similar usefulness ratings across sources despite differential impact. More broadly, we advocate for considering perceptions as theoretically meaningful perspectives in conjunction with more objective evidence of feedback uptake and impact.

5.3.2. When students diverge in how they respond to feedback

Our results suggest significant instructional implications of the moderating role of feedback literacy. Students high in feedback literacy not only perceived teacher feedback more positively in terms of fairness, but also benefited more in the quality of their revised work. This confirms arguments from Carless and Boud [15] and Winstone et al. [16] that feedback literacy should be treated as an essential learning outcome in its own right. Importantly, building feedback literacy is not a passive process but requires structured interventions, such as explicit instruction on how to interpret and apply feedback [65,93], guided peer review activities [11], or reflective exercises that prompt students to evaluate and plan based on received feedback [99]. Embedding these activities across the curriculum—rather than treating feedback as a one-off event—could help cultivate a more feedback-savvy student body capable of benefiting from a broader array of feedback sources, including AI systems whose pedagogical quality may vary.

Student motivation also emerged as a key factor, particularly in relation to the willingness to act upon LLM-generated feedback. Intrinsically motivated students were more likely to engage with and apply feedback from AI, while extrinsic motivation showed no such moderating effects. This finding resonates with self-determination theory [13], which posits that intrinsic motivation enhances deep engagement with learning activities. It also aligns with observations from Naismith and Lajoie [62] that autonomous motivation is particularly critical in self-directed learning contexts where external structures are minimal. In the context of LLM-based feedback, where feedback may lack personalized motivational support and emotional resonance, students' self-motivation becomes crucial for sustained engagement. Therefore, efforts to foster intrinsic motivation—through autonomy-supportive teaching [94], emphasizing relevance and personal meaning in assignments [95], and giving students agency in their learning pathways—may be necessary complements to the implementation of AI-enhanced feedback systems.

Taken together, our findings suggest that carefully combining feedback sources while also laying the groundwork for feedback uptake by supporting students' feedback literacy and intrinsic motivation could be particularly powerful to foster a constructive feedback culture in the classroom. Instead of relying on LLMs as stand-alone solutions, educational designs could combine immediate AI feedback with structured opportunities for reflection, peer discussion, and teacher follow-up to promote deeper engagement and critical evaluation of feedback. Our study suggests that teacher feedback proved most effective for higher-order scientific argumentation, while peer and LLM feedback performed similarly well on formal aspects and were perceived as fair and easy to accept. LLM feedback also offered immediacy and consistency. Rather than suggesting a general "mix of sources," our findings indicate that teacher input is best reserved for cognitively demanding, higher-order aspects, while peer and LLM feedback can effectively support more routine or formal dimensions of writing. Combining these sources therefore entails a division of labor aligned to their comparative strengths (for a comparison of *perceived* strengths of GenAI vs. teacher feedback, [96]). In doing so, educators can leverage the scalability benefits of AI while safeguarding the pedagogical richness that human feedback provides.

6. Limitations

Several limitations should be noted. Although randomization and blinding strengthened internal validity, our modest sample limited statistical power, especially for moderation analyses. Larger studies may yield more precise estimates and detect smaller effects.

Our study took place in a German-speaking educational sciences course with a relatively homogeneous student group. Prior work shows that perceptions of feedback vary with cultural context, discipline, and age [7,97]. Such contextual factors may have shaped how students engaged with teacher, peer, and LLM feedback, limiting direct transfer to other settings (e.g., STEM courses, younger learners, or culturally diverse classrooms).

Blinding reduced bias from source expectations and biases but also lowered ecological validity, as students normally know the feedback source. Likewise, requiring all students to provide feedback controlled for learning-by-giving but stripped peer assessment of a key strength, potentially reducing impact. Future studies should carefully distinguish evaluations of *peer feedback as a source* from *peer feedback as a method*.

We did not systematically analyze feedback quality, which future work could add for a more holistic comparison. Our focus on short-term revisions leaves open questions about long-term effects and transfer. Although randomization reduced confounding by prior attitudes, unmeasured experiences (e.g., familiarity with peer review or AI tools) may still shape engagement. It is also possible that some students inferred the feedback source, which could have influenced perceptions, even though humans are generally unreliable at AI-text detection [98].

Finally, while some of our outcome measures rely on student self-reports, which can diverge from performance-based indicators, our design deliberately combined these subjective and objective measures to capture both perceptions and actual feedback quality.

More generally, research on LLM feedback faces the challenge of rapid system development. While our findings likely generalize to near-term improvements, future models may provide far more effective feedback, requiring ongoing reassessment of how AI, peer, and teacher feedback can best be combined to maximize learning.

7. Conclusions

In an era where scalable feedback solutions are increasingly available—and necessary—this study provides fresh insights into the comparative strengths and limitations of teacher, peer, and AI-generated feedback. While all three sources can contribute meaningfully to student learning, their effectiveness depends on the interplay between task demands, student capabilities, and motivational dispositions. Our findings reinforce that technology can augment, but not replace, the pedagogical expertise of human instructors, particularly when it comes to fostering complex academic skills. This is indicated by our finding that teacher feedback led to significantly better revision quality, compared to peer- and LLM-feedback. Our findings also highlight feedback literacy and intrinsic motivation as crucial ingredients for students to navigate and benefit from diverse feedback landscapes. As AI continues to transform educational practice, the future of feedback lies in designing ecosystems where human judgment, peer collaboration, and artificial intelligence complement one another to support deeper, more autonomous student learning.

Availability of data and material

Data for this research cannot be made available upon request.

A preregistration and supplementary material can be found online at OSF at: <https://doi.org/10.17605/OSF.IO/EZANP>

Ethics declaration

Informed consent for study participation as well as consent to publish the data in an academic paper was obtained from all study participants prior to collecting their data. The university provides a self-assessment form for ethical approval. As this study meets all criteria, no formal approval was required.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI in the writing process

During the preparation of this work the authors used Open AI ChatGPT (GPT-4o for the original submission; GPT-5 for revision) for proofreading and stylistic improvements. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRedit authorship contribution statement

Joshua Weidlich: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Flurin Gotsch:** Writing – review & editing, Software, Investigation, Data curation, Conceptualization. **Kai Schudel:** Writing – review & editing, Project administration, Conceptualization. **Claudia Marusic-Würscher:** Writing – review & editing, Investigation, Data curation, Conceptualization. **Jennifer Mazzarella:** Project

administration, Investigation. **Hannah Bolten:** Writing – review & editing, Investigation, Data curation. **Dario Büttler:** Writing – review & editing, Investigation, Data curation. **Simon Luger:** Writing – review & editing, Investigation. **Bettina Wohlfender:** Investigation. **Katharina Maag Merki:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

APPENDIX A: Preparation and instruction of ChatGPT for LLM feedback

1. Preparation of ChatGPT

Procedure:

1. Set up GPT (formulate instructions, upload knowledge base, test with past assignments; minimal iterations for structural output, none for content).
2. Enter each student assignment individually and anonymized into GPT as plain text.
3. Copy GPT feedback into a standardized Word template.

Model specification:

- OpenAI ChatGPT, model 4.0.
- Internet search enabled.

Knowledge base included:

- Readings required of students.
- Summaries of lecture transcripts (since GPT has input length limits).
 - Transcripts created with OpenAI Whisper base, then summarized by GPT-4, summaries reviewed by the instructor.
- Lecture and exercise slides (PDF).
- Written assignment instructions (subtask 1a).
- Rubric and feedback template.

2. Instructions given to GPT (translated from German language)

[START OF PROMPT]

“Task for GPT: Assessment and feedback generation

GPT’s task is to evaluate student essays and provide structured, actionable feedback in German. Before starting the evaluation, GPT reads the “Assignment 1a” description and all course materials stored in the “Knowledge” area (including lecture slides, exercise slides, and assigned readings). These materials serve as the foundation for evaluation.

Evaluation process and requirements

1. Fill in the rubric

GPT evaluates the essay directly in the rubric (“Feedback-Template”) stored in the “Knowledge” area.

- **Criterion selection:** For each rubric criterion, GPT selects the appropriate performance indicator and marks the category with an “X.” A critical perspective is explicitly desired to give students clear, detailed feedback.
- **Comments:** For at least three criteria with the most room for improvement, GPT writes a short comment (1–2 full sentences). These comments must refer to **concrete examples from the essay** and include **specific suggestions for improvement**. Suggestions should be explicitly linked to course materials in the “Knowledge” base.

2. Write the narrative feedback

After filling in the rubric, GPT produces a final narrative feedback (120–200 words). This text:

- Highlights positive aspects of the essay, with concrete examples.

- Provides constructive, detailed suggestions for improvement, including methods or approaches, linked to course content, and phrased as actionable advice.
3. **Check formal criteria**
Feedback also checks whether the essay meets formal requirements:
- Length: 800–1000 words (excluding references).
 - Font: Times New Roman, 12 pt, 1.5 line spacing.
 - Spelling: orthographically correct, clear German.
 - References: APA 7 style applied correctly.
4. **Content requirements**
Evaluation covers whether the essay addresses the following tasks:
- Positioning of the three quality areas (school quality/ effectiveness framework, with literature).
 - Identification of one development need per area.
 - Identification of one strategy per development need, with theoretical/empirical justification.
 - For each strategy, a limitation or challenge, justified with literature.
5. **Rubric criteria**
1. **Quality areas...**
- ...not identifiable or incorrectly located / not or incorrectly justified with scholarly references
 - ...partially or imprecisely located / weakly justified with scholarly references
 - ...completely and precisely located / correctly justified with scholarly references
2. **Development needs...**
- ...not identified / entirely incorrectly described
 - ...partially correctly identified / partially plausibly described
 - ...correctly identified / fully plausibly described
3. **Action strategies...**
- ...not identified / not or incorrectly justified with scholarly references
 - ...partially identified / weakly justified with scholarly references
 - ...correctly identified / correctly justified with scholarly references
4. **Limitations...**
- ...not identified / not or incorrectly justified with scholarly references
 - ...partially identified / weakly justified with scholarly references
 - ...correctly identified / correctly justified with scholarly references
5. **Argumentative development**
- Statements are largely independent of one another, many breaks in argumentation
 - Statements are partly independent of one another, occasional breaks in argumentation
 - Statements are logically connected, a clear line of reasoning is evident
6. **Orthography**
- Many disruptive errors
 - A moderate number of errors, some disruptive
 - (Nearly) error-free text
7. **Academic language**
- Not neutral / few technical terms / much colloquial language
 - Mostly neutral / only some technical terms / some colloquial language
 - Fully neutral language / consistent use of technical terms / no colloquial language
8. **Paraphrasing, citing APA 7**

- Consistently incorrect paraphrasing and citation
- Partially incorrect paraphrasing and citation
- Consistently correct paraphrasing and citation

9. Formal criteria (length, font, reference list)

- Formal requirements not met
- Formal requirements partially met
- Formal requirements met

Output requirements:

- GPT output should contain only:
- The rubric with judgments,
 - The selected comments on criteria,
 - The final narrative feedback.

Personal address should use informal “Du” when directly addressing the student.”

[END OF PROMPT]

References

- [1] Hattie J, Timperley H. The power of feedback. *Rev Educ Res* 2007;77(1):81–112.
- [2] Wisniewski B, Zierer K, Hattie J. The power of feedback revisited: A meta-analysis of educational feedback research. *Front Psychol* 2020;10:487662.
- [3] Butler DL, Winne PH. Feedback and self-regulated learning: A theoretical synthesis. *Rev Educ Res* 1995;65(3):245–81.
- [4] Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Stud High Educ* 2006;31(2):199–218.
- [5] Boud D, Molloy E. Rethinking models of feedback for learning: the challenge of design. *Assess Eval High Educ* 2013;38(6):698–712.
- [6] Henderson M, Ryan T, Phillips M. The challenges of feedback in higher education. *Assess Eval High Educ* 2019.
- [7] Carless D. Differing perceptions in the feedback process. *Stud high educ* 2006;31(2):219–33.
- [8] Er E, Akçapınar G, Bayazit A, Noroozi O, Banihashem SK. Assessing student perceptions and use of instructor versus AI-generated feedback. *Br J Educ Technol* 2025;56(3):1074–91.
- [9] Jansen T, Höft L, Bahr L, Fleckenstein J, Möller J, Köller O, Meyer J. Comparing generative AI and expert feedback to students’ writing: insights from student teachers. *Psychol Erzieh Unterr* 2024;71(2):80–92.
- [10] Huisman B, Saab N, Van Den Broek P, Van Driel J. The impact of formative peer feedback on higher education students’ academic writing: a meta-analysis. *Assess Eval High Educ* 2019;44(6):863–80.
- [11] Zong Z, Schunn CD, Wang Y. What aspects of online peer feedback robustly predict growth in students’ task performance? *Comput Hum Behav* 2021;124:106924.
- [12] Winstone NE, Hepper EG, Nash RA. Individual differences in self-reported use of assessment feedback: the mediating role of feedback beliefs. *Educ Psychol (L)* 2021;41(7):844–62.
- [13] Deci EL, Ryan RM. The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychol Inq* 2000;11(4):227–68.
- [14] Panadero E, Lipnevich AA. A review of feedback models and typologies: towards an integrative model of feedback elements. *Educ Res Rev* 2022;35:100416.
- [15] Carless D, Boud D. The development of student feedback literacy: enabling uptake of feedback. *Assess Eval High Educ* 2018;43(8):1315–25.
- [16] Winstone NE, Nash RA, Parker M, Rowntree J. Supporting learners’ agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educ Psychol* 2017;52(1):17–37.
- [17] Meyer J, Jansen T, Daumiller M, Fleckenstein J. Understanding individual differences in students’ responses to technology-based feedback on a writing task: the role of achievement motives and initial task performance. *J Res Technol Educ* 2025:1–31.
- [18] Ruwe T, Mayweg-Paus E. Embracing LLM Feedback: the role of feedback providers and provider information for feedback effectiveness. *Frontiers in education*, 9. *Frontiers Media SA*; 2024, 1461362.
- [19] Banihashem SK, Kerman NT, Noroozi O, Moon J, Drachler H. Feedback sources in essay writing: peer-generated or AI-generated feedback? *Int J Educ Technol High Educ* 2024;21(1):23.
- [20] Dai W, Tsai YS, Lin J, Aldino A, Jin H, Li T, Chen G. Assessing the proficiency of large language models in automatic feedback generation: an evaluation study. *Comput Educ: Artif Intell* 2024;7:100299.
- [21] Li K, Lan J, Hu Y. Comparative analysis of GPT-4.0 and teacher feedback on student-generated questions in the flipped classroom. *Educ Technol Res Dev* 2025: 1–23.
- [22] Nazaretsky T, Mejia-Domenzain P, Swamy V, Frej J, Käser T. AI or human? Evaluating student feedback perceptions in higher education. In: *European Conference on Technology Enhanced Learning*. Cham: Springer Nature Switzerland; 2024. p. 284–98.
- [23] Steiss J, Tate T, Graham S, Cruz J, Hebert M, Wang J, Olson CB. Comparing the quality of human and ChatGPT feedback of students’ writing. *Learn Instr* 2024;91:101894.
- [24] Shute VJ. Focus on formative feedback. *Rev Educ Res* 2008;78(1):153–89.

- [25] Winstone NE, Nash RA. Toward a cohesive psychological science of effective feedback. *Educ Psychol* 2023;58(3):111–29.
- [26] Fong CJ, Patall EA, Vasquez AC, Stautberg S. A meta-analysis of negative feedback on intrinsic motivation. *Educ Psychol Rev* 2019;31:121–62.
- [27] Weidlich J, Fink A, Jivet I, Yau J, Giorgashvili T, Drachslers H, Frey A. Emotional and motivational effects of automated and personalized formative feedback: the role of reference frames. *J Comput Assist Learn* 2024;40(6):2735–52.
- [28] Weidlich J, Fink A, Frey A, Jivet I, Gombert S, Menzel L, Drachslers H. Highly informative feedback using learning analytics: how feedback literacy moderates student perceptions of feedback. *Intern J Edu Technol Higher Educ* 2025;22(1):43.
- [29] Dijks MA, Brummer L, Kostons D. The anonymous reviewer: The relationship between perceived expertise and the perceptions of peer feedback in higher education. *Assess Eval High Educ* 2018;43(8):1258–71.
- [30] Ruegg R. Differences in the uptake of peer and teacher feedback. *RELC J* 2015;46(2):131–45.
- [31] Daou D, Sabra R, Zgheib NK. Factors that determine the perceived effectiveness of peer feedback in collaborative learning: A mixed methods design. *Med Sci Educ* 2020;30:1145–56.
- [32] Zhang Y, Schunn CD. Self-regulation of peer feedback quality aspects through different dimensions of experience within prior peer feedback assignments. *Contemp Educ Psychol* 2023;74:102210.
- [33] Liu NF, Carless D. Peer feedback: the learning element of peer assessment. *Teach High Educ* 2006;11(3):279–90.
- [34] Strijbos JW, Narciss S, Dünnebier K. Peer feedback content and sender's competence level in academic writing revision tasks: are they critical for feedback perceptions and efficiency? *Learn Instr* 2010;20(4):291–303.
- [35] Gao Y, Schunn CDD, Yu Q. The alignment of written peer feedback with draft problems and its impact on revision in peer assessment. *Assess Eval High Educ* 2019;44(2):294–308.
- [36] Rotsaert T, Panadero E, Schellens T. Anonymity as an instructional scaffold in peer assessment: its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *Eur J Psychol Educ* 2018;33(1):75–99.
- [37] Van Gennip NA, Segers MS, Tillema HH. Peer assessment for learning from a social perspective: the influence of interpersonal variables and structural features. *Educ Res Rev* 2009;4(1):41–54.
- [38] Van Gennip NA, Segers MS, Tillema HH. Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learn Instr* 2010;20(4):280–90.
- [39] Panadero E, Alqassab M. An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assess Eval High Educ* 2019.
- [40] Nicol D, Thomson A, Breslin C. Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education* 2014;39(1):102–22. <https://doi.org/10.1080/02602938.2013.795518>.
- [41] Deeva G, Bogdanova D, Serral E, Snoeck M, De Weerd T. A review of automated feedback systems for learners: classification framework, challenges and opportunities. *Comput Educ* 2021;162:104094.
- [42] Cavalcanti AP, Barbosa A, Carvalho R, Freitas F, Tsai YS, Gašević D, Mello RF. Automatic feedback in online learning environments: A systematic literature review. *Comput Educ: Artif Intell* 2021;2:100027.
- [43] Keuning H, Jeuring J, Heeren B. A systematic literature review of automated feedback generation for programming exercises. *ACM Trans Comp Educ (TOCE)* 2018;19(1):1–43.
- [44] Pecaric M, Boutis K, Beckstead J, Pusic M. A big data and learning analytics approach to process-level feedback in cognitive simulations. *Acad Med* 2017;92(2):175–84.
- [45] Knight S, Shibani A, Abel S, Gibson A, Ryan P. AcaWriter: A learning analytics tool for formative feedback on academic writing. *J Writ Res* 2020.
- [46] Drachslers H. Towards highly informative learning analytics. Heerlen: Open Universiteit; 2023.
- [47] Gombert S, Fink A, Giorgashvili T, Jivet I, Di Mitri D, Yau J, Drachslers H. From the automated assessment of student essay content to highly informative feedback: A case study. *Int J Artif Intell Educ* 2024;34(4):1378–416.
- [48] Meyer J, Jansen T, Schiller R, Liebenow LW, Steinbach M, Horbach A, Fleckenstein J. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Comp Edu: Artif Intel* 2024;6:100199.
- [49] Zhang DW, Boey M, Tan YY, Jia AHS. Evaluating large language models for criterion-based grading from agreement to consistency. *npj Sci Learn* 2024;9(1):79.
- [50] Richter E, Spitzer MWH, Morgan A, Frede L, Weidlich J, Moeller K. Large language models outperform humans in identifying neuromyths but show sycophantic behavior in applied contexts. *Trends Neurosci Educ* 2025:100255.
- [51] Sun, Y., & Wang, T. (2025). Be friendly, not friends: how LLM sycophancy shapes user trust. *arXiv preprint arXiv:2502.10844*.
- [52] Usher M. Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education. *Assess Eval High Educ* 2025:1–16.
- [53] Brooks C, Burton R, Van der Kleij F, Ablaza C, Carroll A, Hattie J, Salinas JG. "It actually helped": students' perceptions of feedback helpfulness prior to and following a teacher professional learning intervention. In: *Frontiers in Education*. (Vol. 9). Frontiers Media SA; 2024. p. 1433184.
- [54] Mandouit L, Hattie J. Revisiting "The Power of Feedback" from the perspective of the learner. *Learn and Instr* 2023;84:101718.
- [55] Brenner PS, DeLamater J. Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias. *Soc psyc quar* 2016;79(4):333–54.
- [56] Pekrun R. Self-report is indispensable to assess students' learning. *Front lear res* 2020;8(3):185–93.
- [57] Panadero E. Toward a paradigm shift in feedback research: five further steps influenced by self-regulated learning theory. *Educ Psychol* 2023;58(3):193–204.
- [58] Winstone NE, Nash RA. Developing students' proactive engagement with feedback. *Innovative assessment in higher education*. Routledge; 2019. p. 129–38.
- [59] Howard JL, Bureau JS, Guay F, Chong JX, Ryan RM. Student motivation and associated outcomes: A meta-analysis from self-determination theory. *Perspect Psychol Sci* 2021;16(6):1300–23.
- [60] Park S, Yun H. The influence of motivational regulation strategies on online students' behavioral, emotional, and cognitive engagement. *Am J Distance Educ* 2018;32(1):43–56.
- [61] Wolters CA, Benzon MB. Assessing and predicting college students' use of strategies for the self-regulation of motivation. *J Exp Educ* 2013;81(2):199–221.
- [62] Naismith LM, Lajoie SP. Motivation and emotion predict medical students' attention to computer-based feedback. *Adv Health Sci Educ* 2018;23:465–85.
- [63] Vansteenkiste M, Lens W, Deci EL. Intrinsic versus extrinsic goal contents in self-determination theory: another look at the quality of academic motivation. *Educ Psychol* 2006;41(1):19–31.
- [64] Deslauriers L, McCarty LS, Miller K, Callaghan K, Kestin G. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proc Natl Acad Sci* 2019;116(39):19251–7.
- [65] Molloy E, Boud D, Henderson M. Developing a learning-centred framework for feedback literacy. *Assess Eval High Educ* 2020;45(4):527–40.
- [66] Weidlich J, Jivet I, Woitt S, Orhan Gökşin D, Kraus J, Drachslers H. The student feedback literacy instrument (SFLI): multilingual validation and introduction of a short-form version. *Assess Eval High Educ* 2025:1–17.
- [67] Double KS, McGrane JA, Hopfenbeck TN. The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educ Psychol Rev* 2020;32(2):481–509.
- [68] Motz BA, Carvalho PF, de Leeuw JR, Goldstone RL. Embedding experiments: staking causal inference in authentic educational contexts. *J Learn Anal* 2018;5(2):47–59.
- [69] Weidlich J, Hicks B, Drachslers H. Causal reasoning with causal graphs in educational technology research. *Educ tech res develop* 2024;72(5):2499–517.
- [70] Ross SM, Morrison GR, Lowther DL. Educational technology research past and present: balancing rigor and relevance to impact school learning. *Contemp Educ Technol* 2010;1(1):17–35.
- [71] Kerman WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol* 1999;52(1):19–26.
- [72] Lundstrom K, Baker W. To give is better than to receive: the benefits of peer review to the reviewer's own writing. *J Second Lang Writ* 2009;18(1):30–43.
- [73] Maag Merki K. Schulqualitätsforschung. *Handbuch schulforschung*. Wiesbaden: Springer Fachmedien Wiesbaden; 2021. p. 1–22.
- [74] Merki KM, Wulschleger A, Rechsteiner B. Ein neuer blick auf Schulentwicklung. Das zusammenspiel zwischen impliziten und expliziten prozessen der weiterentwicklung der einzelschule. *Schulentwickl als Theor: Forschungsperspektiven auf Veränderungsprozesse von Sch* 2021:159–80.
- [75] Black P, William D. Developing the theory of formative assessment. *Educ Assess Eval Account (former: J pers eval educ)* 2009;21:5–31.
- [76] Dawson P, Henderson M, Mahoney P, Phillips M, Ryan T, Boud D, Molloy E. What makes for effective feedback: staff and student perspectives. *Assess Eval High Educ* 2019;44(1):25–36.
- [77] Thomas AE, Müller FH, Bieg S. Entwicklung und validierung der skalen zur motivationalen regulation beim lernen im studium (SMR-LS). *Diagnostica* 2018.
- [78] Woitt S, Weidlich J, Jivet I, Orhan Gökşin D, Drachslers H, Kalz M. Students' feedback literacy in higher education: an initial scale validation study. *Teach High Educ* 2025;30(1):257–76.
- [79] Dawson P, Yan Z, Lipnevich A, Tai J, Boud D, Mahoney P. Measuring what learners do in feedback: the feedback literacy behavioural scale. *Assess Eval High Educ* 2024;49(3):348–62.
- [80] Strijbos JW, Pat-El R, Narciss S. Structural validity and invariance of the feedback perceptions questionnaire. *Stud Educ Eval* 2021;68:100980.
- [81] Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. *J Pers Assess* 2003;80(3):217–22.
- [82] Toulmin S. *The Uses of Argument*. 2nd ed. Cambridge University Press; 2003.
- [83] McNeill K, Krajcik J. Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing. Pearson Allyn & Bacon; 2012.
- [84] Cohen J. The effect size. *statistical power analysis for the behavioral sciences*. Abingdon: Routledge; 1988. p. 77–83.
- [85] Nelson MM, Schunn CD. The nature of feedback: how different types of peer feedback affect writing performance. *Instr Sci* 2009;37(4):375–401.
- [86] Fanous A, Goldberg J, Agarwal, A.A., Lin, J., Zhou, A., Daneshjouri, R., & Koyejo, S. (2025). SycEval: evaluating LLM sycophancy. *arXiv preprint arXiv:2502.08177*.
- [87] Evans C. Making sense of assessment feedback in higher education. *Rev Educ Res* 2013;83(1):70–120.
- [88] Molenaar I. Towards hybrid human-AI learning technologies. *Europ J Edu* 2022;57(4):632–45.
- [89] Sperber, L., MacArthur, M., Minnillo, S., Stillman, N., & Whithaus, C. (2025). Peer and AI Review+ Reflection (PAIRR): A human-centered approach to formative assessment. *Comp and Comp*, 76, 102921.
- [90] Bjork RA, Dunlosky J, Kornell N. Self-regulated learning: beliefs, techniques, and illusions. *Annu Rev Psychol* 2013;64(1):417–44.
- [91] Sitzmann T, Ely K, Brown KG, Bauer KN. Self-assessment of knowledge: A cognitive learning or affective measure?. *Acad Manag Learn Educ*, 9; 2010. p. 169–91.

- [92] Noroozi O, Alqassab M, Taghizadeh Kerman N, Banihashem SK, Panadero E. Does perception mean learning? Insights from an online peer feedback setting. *Assess Eval High Educ* 2025;50(1):83–97.
- [93] Sutton P. Conceptualizing feedback literacy: knowing, being, and acting. *Innov Educ Teach Int* 2012;49(1):31–40.
- [94] Reeve J. Teachers as facilitators: what autonomy-supportive teachers do and why their students benefit. *Elem Sch J* 2006;106(3):225–36.
- [95] Niemiec CP, Ryan RM. Autonomy, competence, and relatedness in the classroom: applying self-determination theory to educational practice. *Theory Res Educ* 2009; 7(2):133–44.
- [96] Henderson M, Bearman M, Chung J, Fawns T, Buckingham Shum S, Matthews KE, de Mello Heredia J. Comparing generative AI and teacher feedback: student perceptions of usefulness and trustworthiness. *Assess Eval High Educ* 2025.
- [97] Nicol D. From monologue to dialogue: improving written feedback processes in mass higher education. *Assess & Eval High Education* 2010;35(5):501–17.
- [98] Chein J, Martinez S, Barone A. Can human intelligence safeguard against artificial intelligence? Exploring individual differences in the discernment of human from AI texts. *Res Sq* 2024. rs-3.
- [99] Winstone N, Carless D. Designing effective feedback processes in higher education: A learning-focused approach. Routledge; 2019.