
CAN WE TRUST LLMs AS A TUTOR FOR OUR STUDENTS?

EVALUATING THE QUALITY OF LLM-GENERATED FEEDBACK IN STATISTICS EXAMS

Markus Herklotz*

Social Data Science and AI Lab

LMU Munich

Munich, Germany

**Corresponding author*

m.herklotz@lmu.de

Niklas Ippisch

Social Data Science and AI Lab

LMU Munich

Munich, Germany

Anna-Carolina Haensch

Social Data Science and AI Lab

LMU Munich

Munich, Germany

November 7, 2025

ABSTRACT

One of the central challenges for instructors is offering meaningful individual feedback, especially in large courses. Faced with limited time and resources, educators are often forced to rely on generalized feedback, even when more personalized support would be pedagogically valuable. To overcome this limitation, one potential technical solution is to utilize large language models (LLMs). For an exploratory study using a new platform connected with LLMs, we conducted a LLM-corrected mock exam during the "Introduction to Statistics" lecture at the University of Munich (Germany). The online platform allows instructors to upload exercises along with the correct solutions. Students complete these exercises and receive overall feedback on their results, as well as individualized feedback generated by GPT-4 based on the correct answers provided by the lecturers. The resulting dataset comprised task-level information for all participating students, including individual responses and the corresponding LLM-generated feedback. Our systematic analysis revealed that approximately 7 % of the 2,389 feedback instances contained errors, ranging from minor technical inaccuracies to conceptually misleading explanations. Further, using a combined feedback framework approach, we found that the feedback predominantly focused on explaining why an answer was correct or incorrect, with fewer instances providing deeper conceptual insights, learning strategies or self-regulatory advice. These findings highlight both the potential and the limitations of deploying LLMs as scalable

feedback tools in higher education, emphasizing the need for careful quality monitoring and prompt design to maximize their pedagogical value.

Keywords: Automated Feedback, Individual Feedback, LLMs, Learning Analytics, Higher Education

1 Introduction

Feedback can be a highly powerful educational tool (Wisniewski et al., 2020). However, delivering individualized feedback in large-scale courses with several hundred students presents a considerable challenge for instructors (Topali et al., 2024). Due to time and resource constraints, instructors might not be able to deliver personalized feedback on their own at all, resorting to whole-class comments or methods such as peer-assessment (Sun et al., 2014) instead. This means many students in high-enrollment classes may not receive the individual feedback by their instructors that they need to improve. One promising technical approach to address this limitation is the integration of Large Language Models (LLMs) as virtual tutors (Liu et al., 2025), capable of generating personalized feedback at scale. However, despite their potential, LLM tutors might exhibit notable weaknesses, such as the generation of erroneous feedback (Jia et al., 2024) and a lack of pedagogical nuance without careful design (Dai et al., 2023) .

A central consideration for instructors to implement automated technologies in their teaching is the *trustworthiness* of these systems for direct student interaction (Feldman-Maggor et al. (2025), Nazaretsky et al. (2022), Viberg et al. (2024), Ayanwale et al. (2024), Lyu et al. (2025)). Trust, in this context, can be understood as an instructor’s willingness to rely on the system and accept its accompanying vulnerabilities, grounded in the belief that generative AI can reliably enhance educational outcomes (Lyu et al., 2025, p. 2). Conversely, distrust arises when such systems are perceived as *unreliable, harmful, or misaligned with instructional goals*, prompting educators to withhold reliance or limit use (Lyu et al., 2025, p. 2).

In order to leverage the capabilities of LLMs for generating individualized student feedback while minimizing errors, we used the platform *StudyLabs*, on which students received automated, personalized feedback generated by the LLM GPT-4 based on correct solutions provided by instructors (fig. 1). To generate the feedback, *StudyLabs* submits the students’ solutions, scoring guidelines, the correct answers, and the original task descriptions to the OpenAI API. This integration aims to ensure that the feedback provided is consistent with the instructor’s expectations and aligned with the intended learning outcomes of the course.

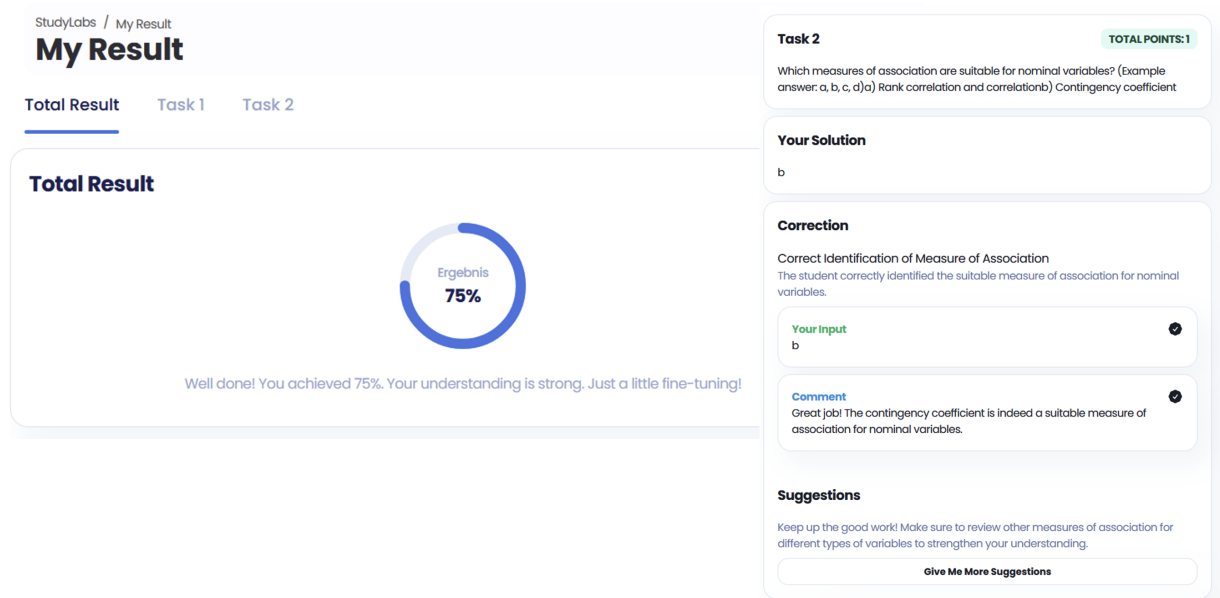


Figure 1: Screenshots of receiving feedback on the *StudyLabs* platform.

Nonetheless, there remains a risk that LLMs may produce misleading, incomplete, or inconsistently structured feedback (Li et al., 2023). Prior research has shown that LLM-generated feedback can still contain substantial errors and hallucinations, even when given correct solutions (Jia et al., 2024). Hence, it is necessary to determine whether, and to what extent, the feedback produced under these conditions still contains factual inaccuracies or misleading answers. In addition, reliability extends beyond factual correctness to the structure and consistency of feedback: students who provide identical answers should potentially receive comparable guidance, yet LLMs may respond differently to each case (Jacobsen and Weber, 2023; Liang et al., 2024), potentially creating unequal learning opportunities. Instructors need a clear understanding of the nature, structure, and quality of the feedback that will be provided to their students.

In this study, we systematically analyze LLM-generated feedback from a mock exam in an introductory statistics lecture at a German university. We first identify and classify errors to assess their prevalence and discuss their potential pedagogical impact. We then evaluate the feedback based on established theory, extending the Hattie and Timperley (2007) feedback model by subdividing *task-level* comments into *right/wrong*, *response-oriented*, and *conceptually-focused* (Ryan et al., 2020), while retaining the original *process*, *self-regulatory*, and *self* levels. We examine how the feedback was usually structured, its variance by task type and performance level and whether students with identical scores receive comparable guidance. Across 2,389 GPT-4 feedback instances from an authentic classroom setting, we quantify error prevalence and propose a combined feedback framework that supports our structured analysis and opens new avenues of designing feedback for future task-based learning environments.

In the *Background* section, we first discuss feedback theories in general before addressing recent studies specifically analyzing LLM-generated feedback. Based on this, we then present our assessment framework and research questions. Following this, we present our methodology, detailing the sample characteristics and the instruments employed for

data collection and analysis. In the subsequent *Results* section, we offer a comprehensive analysis of the feedback, incorporating both quantitative and qualitative perspectives. Finally, in the *Discussion and Conclusion* section, we contextualize these findings, draw conclusions, and outline directions for future research.

2 Background

2.1 What is Feedback?

If used appropriately, feedback can be a highly powerful educational tool (Wisniewski et al., 2020). Accordingly, many studies have investigated what constitutes effective feedback and under which circumstances (Lipnevich and Panadero, 2021). In their meta-review, Lipnevich and Panadero (2021) analyzed 14 different feedback models and reported that Hattie and Timperley (2007) was both the (by far) most-cited (>14,000 citations at the time) and the only framework unanimously selected by their consulted.

Following this popular model, feedback can be ‘conceptualized as information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one’s performance or understanding’ (Hattie and Timperley, 2007, p. 81).

Effective feedback addresses three major questions: *Where am I going? How am I going? Where to next?* When there is a discrepancy between what a student understands and what is aimed to be understood, the answers to these feedback questions can enhance learning. In these instances, feedback can help to reduce this discrepancy by raising motivation and facilitating processes that lead to understanding. The model outlined by Hattie and Timperley (2007) differentiates between four levels of feedback that influence its effectiveness:

- *Task level*: How well a task is performed
- *Process level*: Strategies for the process of solving the task
- *Self-regulatory level*: Strategies for self-evaluation and engagement based on already existing knowledge
- *Self level* : Feedback directed to the recipient on a personal level (usually praise)

While the self level in Hattie and Timperley (2007) is considered to be rarely effective, the other three levels are interrelated. Effective feedback moves from the task to the process to the self-regulatory level. This ensures that the focus is not on the specifics of the task itself but leads instead to higher engagement with and deeper understanding of the underlying concepts.

Extending this model, recent work of Hattie from the learner’s perspective emphasizes that feedback will only be used when it is heard, understandable, and actionable, and when students judge the “transaction costs” of engaging with it (effort, time, emotional risk) to be worth the anticipated performance gains (Mandouit and Hattie, 2023). Purely past-oriented comments or correctness checks therefore have limited educative value unless they include explicit feed-forward and built-in opportunities to apply guidance in subsequent tasks (Mandouit and Hattie, 2023). Ideally this is part of an ongoing dialog among students, peers, and teachers that develops feedback literacy, shifting the criterion of

effective feedback from what is given to what is actually understood and acted upon by learners (Mandouit and Hattie, 2023).

2.2 Research on LLM feedback

In recent years, there have been many studies that evaluated LLM-generated feedback, following different approaches and use cases. One main difference in those studies are the various types of exercises they use for their feedback implementations. We can observe two predominant task domains for implementation: elaborative writing tasks such as essays (Dai et al., 2023; Gombert et al., 2024; Meyer et al., 2024; Steiss et al., 2024; Jansen et al., 2024; Gombert et al., 2024; Venter et al., 2025; Wan and Chen, 2024) and programming exercises (Estevez-Ayres et al., 2024; Phung et al., 2023, 2024; Roest et al., 2023; Hellas et al., 2023; Jia et al., 2024; Lohr et al., 2025; Er et al., 2025).

While these different types of exercises are quite obvious to classify, the underlying theoretical frameworks and evaluation schemes for assessing LLM-generated feedback remain highly fragmented. Some studies purely focused on the assessment of errors and technical accuracy (Estevez-Ayres et al., 2024; Jia et al., 2024), whereas others prioritize the pedagogical quality and perceived usefulness of the feedback (Jansen et al., 2024; Meyer et al., 2024; Er et al., 2025; Gombert et al., 2024). Another line of research integrates both dimensions, evaluating error occurrence together with quality criteria often comparing LLM-generated with human feedback (Dai et al., 2023; Lohr et al., 2025; Jürgensmeier and Skiera, 2024; Wan and Chen, 2024). Just as the methodological approaches vary widely, so too do the findings reported across the studies. On the one hand, there is evidence that LLM feedback might increase students' performance compared to receiving no feedback (Makransky et al., 2025) and was rated positively by students regarding the tone and scope (Roest et al., 2023; Jürgensmeier and Skiera, 2024). At the same time, students criticize that LLM feedback lacks depth and helpfulness (Jürgensmeier and Skiera, 2024; Roest et al., 2023) and preferred human feedback over LLM feedback (Jansen et al., 2024). Also in terms of students' performance, there is evidence that human feedback still outperforms LLM feedback (Makransky et al., 2025). When analyzed by experts, the evidence is contradictory: some find that LLM feedback outperforms human feedback, e.g., in terms of correctness, informativeness, conciseness, and comprehensibility (Phung et al., 2023, 2024), or in terms of mentioning the relevant criteria, specificity, and mentioning additional explanations (Jacobsen and Weber, 2023). Other studies found that LLM feedback is not able to outperform human feedback, for example when assessing if the feedback provided instructions for improvement, was accurate, prioritized essential features, and used a supportive tone (Steiss et al., 2023). In terms of errors, some studies raise concerns: LLM feedback is often not able to detect the issue (Estevez-Ayres et al., 2024; Hellas et al., 2023) or reports non-existent issues (Hellas et al., 2023). Experts assessing the feedback in the study by Wan and Chen (2024) found that one third of feedback needs major revisions or needs to be rewritten completely. Overall, the previous research indicates that LLM feedback performs well to some extent, especially when it comes to the style and tone, and might be preferred over no feedback. On the other hand, these new research approaches also suggest moderate to high prevalence of errors as well as concerns whether it can outperform human feedback.

2.3 A Combined Feedback Framework

Our central question is whether teachers can rely on LLMs with providing feedback to their students. Hence, we must weigh accuracy and instructional value in equal measure. The relevant counterfactual in a large-enrollment course is no individual feedback at all, so we will not benchmark against human feedback. Instead, we judge the LLM output against established feedback theory, most prominently the four-level model of Hattie and Timperley (2007), still the field’s dominant framework (Lipnevich and Panadero, 2021).

Empirical work shows a persistent skew toward the task level in both human and AI feedback (for example recently by Dai et al. (2023)). Process, self-regulatory and self feedback typically account for a very small amount of feedback. Given this consistent trend, we anticipate our LLM to produce predominantly task-level comments as well. To yield deeper analytic and qualitative insight, we further dissect the task-level category. For this purpose, we draw on the feedback taxonomy by (Ryan et al., 2020), which we expect to comprehensively capture the spectrum of task-level feedback, especially for our statistics mock exam:

1. *Right/wrong feedback*: Simple identification of the correct response
2. *Response-oriented feedback*: Short explanation of why each option was correct or incorrect
3. *Conceptually-focused feedback*: Discussion of the correct response in more detail (e.g., the underlying concepts)

In their study, (Ryan et al., 2020) employed different tasks related to each other to analyze the effect of these kinds of feedback on the near- and far-transfer of knowledge. They found "response-oriented" feedback and "conceptually-focused" feedback superior to simple "right/wrong" feedback.

For the evaluation of the LLM generated feedback, we will combine the approaches by Hattie and Timperley (2007) and Ryan et al. (2020). We will dissect the task level in the three categories: "right/wrong," "response-oriented," and "conceptually-focused" based on Ryan et al. (2020). This refinement also aligns with more recent developments in the Hattie framework, which from a learner’s perspective emphasizes that purely corrective comments have limited educative value (Mandouit and Hattie, 2023, see Section 2.1). Accordingly, contrasting the three task-level categories from simple "right/wrong" comments to future-oriented "conceptually-oriented" guidance should yield clearer distinctions and finer-grained insights regarding into the structure and pedagogical value of the feedback. The other Hattie and Timperley (2007) levels remain as they are. Accordingly, we define the feedback categories and their guiding questions for our analysis based on the classifications as follows:

1. **Task level:**
 - (a) **Right/wrong**: Does the feedback text indicate whether the answer is correct or not?
 - (b) **Response-oriented**: Does the feedback text explain why the answer is correct or not?
 - (c) **Conceptually-focused**: Does the feedback text provide an explanation that enhances understanding of the underlying concept of the task?

2. **Process level:** Does the feedback text provide strategies to facilitate the student's learning?
3. **Self-Regulatory level:** Does the feedback text provide suggestions on how the student can control and manage their own learning?
4. **Self level:** Does the feedback text evaluate the student on a personal level?

2.4 Research Questions

The present study investigates the quality and structure of feedback generated by LLMs in educational contexts. Specifically, it examines the occurrence of errors in feedback when correct solutions are available, as well as the structural characteristics of the feedback produced. The analysis is guided by two overarching research questions: (1) To what extent does the LLM produce errors while giving feedback when the instructor provides the correct answers? and (2) What is the structure of the feedback generated, including its variations across tasks and students, and its alignment with established educational feedback theories?

3 Data & Methodology

In our analysis, we rely on three instruments: Data gathered from an online survey both at the start (1) and at the end (2) of the course, as well as the passively tracked data from the mock exam on *StudyLabs* (3). The sample and the methods used for the analysis will be outlined below.

3.1 Sample of students

Our study subjects are Bachelor students who participated in the 'Introduction to Statistics' lecture at the LMU Munich during the winter term 2024. This is a large course of about 300 students¹, of whom 202 completed the entry survey in the first week of the term. The course is mandatory either for students with a major in sociology or (media) informatics, or for students with 'Statistics and Data science' as their minor. The lecture was held in person and streamed online; both the weekly exercise and tutorial sessions were conducted in person only. The course was held entirely in German. Both surveys were programmed online but fielded in person during the lecture to raise response rates. The entry survey reveals that most students in this course are enrolled as Bachelor Sociology majors (75.12%). Among the other majors, Media Informatics (10.4%) and Geography (3.5%) are the most prominent. Regarding the minor study subjects of the students, the biggest group is students with Statistics as their minor (19.7%). The students in this course primarily identify as female (68.3%) and very early in their studies, mainly in their first term (85.6%); some are in the third semester (12.4%).

¹This is about the number of students who take the exam every year. In-class participation is not mandatory. Therefore, it is not possible to provide exact numbers of participants.

3.2 Mock exam and student survey

For gathering insight on the feedback given by the LLM on the *StudyLabs* platform, one of the last sessions of the course in January 2024 was used to host a mock exam in person. The mock exam was used one year prior as the actual exam and was conducted in German. The students took the mock exam on *StudyLabs*² entirely and received the LLM-generated feedback after requesting it (usually after the last exam question). From this mock exam on *StudyLabs*, ZAVI shared the data with the team of authors. The data set contained all questions asked in the mock exam, the achieved and achievable points per exercise, the student answers, and the LLM-generated feedback for all 38 exercises anonymized with a randomly generated username.

In total, we have data from 70 students and their mock exams. Most of the students worked through all exercises, yielding a total of 2,389 exercises and feedback instances. To structure the exam for the analysis, we categorized the 38 exercises into four different types of tasks (knowledge, interpretation, calculation, and R) and eight different categories of statistical concepts (measures of central tendency, measures of dispersion, measures of correlation, visualization, regression, variables, frequencies, and R).³

The in-depth evaluation and coding of the feedback was done by two of the authors. First, one author that is also responsible for the grading of the actual exam of the course, screened all of the 2,389 feedback instances individually for potential errors, yielding a binary variable indicating an error or not. At the same time, we stored the feedback instances with the visually marked errors separately for content analysis. Here, we chose a rather broad definition of error and decided to flag the feedback that might be misleading or confusing for students in their first semester and first statistics course (see also section 4.2). The flagged errors were subsequently checked by another member of the author team. Eight differences between the authors arose, were discussed and resolved, yielding five changes in the flagging.

To assess the feedback structure in detail based on our feedback categories (section 2.3), we chose a subset of the 38 mock exam tasks to be analyzed with a qualitative content analysis. We selected one question from each task category (knowledge/calculation/interpretation) to cover the didactic range of exercises. To allow comparisons within the same task category, we additionally chose a second interpretation question for a total of four exam tasks to be analyzed in detail.

The first question is a multiple choice *knowledge* question, where students were expected to choose measures of central tendency (a) Mean, b) Median, c) Variance, d) Mode). All participating students except one provided an answer and received corresponding feedback (n = 69). Secondly, we chose an *interpretation* question where students were supposed to argue, based on a histogram and boxplot, whether they expected the mean or the median to be higher. All participating

²We deployed *StudyLabs* using version GPT-4-0613 without any changes to the *StudyLabs* system prompt to achieve an authentic classroom setting: "As an AI, you are designed to support students with their exam questions, based on a collection of corrected tasks. Your role is to provide clear, supportive, and constructive answers by integrating the context of the task, model solutions, student responses, assessment criteria, and previous feedback. Use the chat history to maintain continuity in the dialogue and ensure that your responses dynamically align with the evolving conversation. Present mathematical content in KaTeX format and consistently use positive, motivating language. Tailor your explanations to the student's level of understanding and use relevant examples or analogies to clarify complex concepts. Your interaction should be professional, pedagogically sound, and strictly focused on the provided material and student inputs."

³The english translation of the exam including the categorization in task types is available as supplementary material.

students received feedback here ($n = 70$). Lastly, feedback on a *calculation* exercise was analyzed ($n = 69$). Here, variances of weight based on five measurements separately for two groups needed to be calculated.

As a method, a qualitative content analysis (Mayring, 2014) was chosen, in which the feedback categories outlined above were used as categories for coding. The coding was conducted by two of the authors. Before the analysis, a part of one of the tasks was coded independently and then discussed, both to test the coding scheme and to establish inter-coding reliability. Afterwards, the authors split the tasks in half (even and uneven student ID numbers). After the procedure, both authors checked the other persons' coding for possible disagreements, but no were identified. Every feedback text part was coded only with one category. All text passages were coded. Hence, adding up all coded passages yield exactly 100% of feedback, allowing a comparison of the coverage of different categories. Since the provided feedback was in German, we translated examples for the purpose of this paper.

To capture the students' experience, an online survey was designed for the start and the end of the course. The survey was conducted using in-class time to raise participation. Afterwards, the link was also shared on the online learning platform moodle available for everyone that could not attend class. The survey mainly consisted of questions addressing the students' socio-demographic information, previous usage of LLMs and experience with the *StudyLabs* platform.

4 Results

4.1 Students' experience

In the context of introducing an AI tool for students, their experience and perspective is an important part for evaluating the tool. Additional to the evaluation of the *StudyLabs* platform, we asked for the existing experience with ChatGPT⁴. The majority of respondents of the start-of-semester survey ($N=202$) stated that they have used ChatGPT already (54%), and 5% stated that they do not know what ChatGPT is. Of those who said to have used it before, the frequency of usage is very heterogeneous. Regarding the range of use, 43% stated that they use it for explanations, 34% for the production of text, and 23% for programming (multiple answers possible). In the end-of-semester survey ($N = 104$), 65.8% stated that they participated in the mock exam with *StudyLabs*. Among those, most have used ChatGPT already before the mock exam (67 %) with approximately three-quarters of them very frequently.

Of those who stated to have participated at the mock exam and to have used ChatGPT before ($n = 32$), the usage of ChatGPT for explanations (97 %) was most prominent, whereas usage for production of text and programming (both 34 %) was used less (multiple answers possible).

Overall, the feedback comments of the LLM provided on the *StudyLabs*-Platform were rated useful (93 % rather or very useful), and respondents would recommend it (78 % would definitely or rather recommend). One reason for not recommending might be the interface for submitting solutions, which was rated rather low (37 % stated that it was

⁴(Chat)GPT is not the only LLM which applies here, but for example due to media coverage, the most likely for students to have had contact with. We decided that a question detailing other LLMs could have added more confusion among students than the potential benefit of insight.

(rather) not user-friendly.⁵ Interestingly, no difference in the perceived usefulness was found based on whether students already had used ChatGPT before or how frequently they used it.

4.2 Errors in the LLM-generated feedback

Out of the 2,389 feedback instances screened by two coders that are also authors of the paper, 6.99% ($n = 167$) were identified to contain at least one error (see table 1). The errors occurred mainly in feedback for incorrect student answers ($n = 137$) but to a lesser extent also for correct student answers ($n = 30$).

Feedback	Student Answers		Total
	Correct	Incorrect	
No errors	46.92%	46.09%	93.01%
Errors	1.26%	5.73%	6.99%
Total	48.18%	51.82%	2,389

Table 1: Feedback instances containing (no) errors by correct and incorrect student answers.

We observed erroneous feedback instances across various topics and task types in the mock exam, without a consistent pattern (see Table 2). For 13 out of the 38 tasks, the LLM produced only correct responses. The majority of tasks (19) contained between 1 and 5 erroneous feedback instances.⁶ Four tasks resulted in 6 to 10 erroneous feedback instances. Notably, both tasks with 11 to 15 erroneous feedback instances were correlation-related, but drawn from different sections of the exam and pertaining different task types. Two tasks, covering different topics and task types, stood out with much higher error frequencies of over 30 erroneous feedback instances.

These erroneous feedback instances range from individual words being wrong to technical difficulties accepting correct answers to plain wrong explanations of statistical concepts. Instead of categorizing these errors in a broader typology and describing them abstractly, we will first examine them in more qualitative detail in the following section. This detail seems necessary to generate insight into *what* can go wrong and to conceptualize approaches for further reducing errors in the future (see Table 5 in appendix with most common errors and examples).

One example of a technical mistake pertains to question (7a), where students have to interpret a regression coefficient of 0.3021 for the relation between household income and vacation spending. In 10 feedback instances, students answered correctly that this means that per additional Euro of household income, people spend on average *0.3021 Euro* more on vacations. But in these cases, the LLM-generated feedback stated that this is incorrect and that the actual number is *30 cent*. This is very likely due to the provided template solution also noting "30 cents", but neither in the task itself nor in the solution was the criteria for transferring the unit defined.

⁵One potential reason for this rating could be that students were asked to use LaTeX code for submitting formulas, but they were not very familiar with LaTeX yet.

⁶Erroneous feedback instances are responses that contain at least one error, without counting multiple errors within a single response more than once.

Table 2: Error distribution over topics and task types.

Topics	Task types			
	Knowledge	Calculation	Interpretation	R
Central Tendency	1a	3b	2b	
	1h	3c		
	3f	3d		
		3e		
Dispersion	1b	4a		
		4b		
Correlation	1c	5a	5b	
	1e	6b	5c	
		6c	6c	
		6d	7d	
Visualization	1d		2a	
			2c	
			2d	
			2e	
Regression	1f	7b	7a	
	1g	7c	7d	
			7e	
			7f	
			7g	
			7h	
Variables		3a		
Frequencies		6a		
R				2f 4c

Number of erroneous feedback instances	0	1–5	6–10	11–15	31	34
--	---	-----	------	-------	----	----

The most common error was a systematic mistake, which occurred in almost half of the feedback instances ($n = 34$). For this task (3f), where students had to enter the values of the 25 and 75% quartile, a new table was introduced in the previous exercise (3e), but the LLM still referred to the information given at the start of this set of exercises (3a - 3d). Hence, it returned plain wrong information when assigning the quartiles. Another example of not using the correct input was in subsequent tasks where the question description provided an intermediate value in case the student did not solve the previous question. In several instances, the LLM ignored the possibility of this provided intermediate value and just assumed the student's answer was incorrect.

In another common feedback error ($n = 30$), the LLM assumed the wrong scale for the independent variable in a regression. In this task (7h), the variable 'children in the household' was defined as binary (1 = yes, one or more

children, 0 = no children). When returning feedback, the LLM treated this variable as a metric instead of the number of children in a household, resulting in wrong information.

Another confusion occurred repeatedly ($n = 14$) in a question (6b) where students had to calculate χ^2 . The culprits here are the two different, valid approaches to calculating χ^2 : Either the 'traditional' way with first calculating the expected values *or* the faster formula, in which you do not have first to calculate expected values. Both approaches were introduced in the lecture in this course, but the formula without expected values was recommended. Hence, the template solution for this mock exam task also included this faster formula. When the LLM returned feedback for this task, it also showed exactly this formula from the template solution but simultaneously explained the method by first calculating expected values instead. Subsequently, students received an explanation of a formula different from the one shown.

Some other common errors were wrong or insufficient explanations of statistical concepts. This comprises six feedback instances when the LLM only differed between categorical and metric variable scales, using the Spearman and Pearson coefficients evenly while sparing out the information that Spearman is a valid measure for ordinal variables. Further, eight feedback instances showed misleading explanations of an ordinal variable, confusing it with a purely nominal variable.

Further, we also found some minor issues, which are quite heterogeneous and deserve a mention to show the range of feedback errors. In very few (less than five) instances, there were some hallucinations. For example, in a question with a contingency table on the ownership of dogs or cats, the LLM invented birds and fishes in its feedback. In another question, it completely invented a sentence, which we could not decipher where it came from or what its meaning is supposed to be.⁷ As this mock exam and its feedback was conducted in German, in several instances, the translation seemed to go wrong in parts on individual words⁸ or it just used the English word entirely⁹ instead of the German word.

As introduced before, these errors are quite heterogeneous in their types. But they might be distinguishable in two categories based on their potential effect on the students: 1. (More) easily identifiable errors, which mainly cause frustration or distraction, and 2. hardly identifiable errors with convincingly formulated but wrong explanations, which 'silently' introduce incorrect concepts into the students' minds.

In category 1 are the rather technical mistakes, such as not accepting the currency (0.30 Euro vs. 30 cents), ignoring the input of the tasks and the translation, or single word mistakes. Most students should notice quickly that they provided the correct answer when they entered 0.30€ instead of 30 cents. But first of all, it is a frustrating experience to provide the correct answer and not receive the correct grade. Second of all, it could still be misleading for students who were insecure about solving the exercise. While one could argue that the instances of single-word mistakes and translations are minor and should be easily discernible for students, they still can be a distraction in something that is supposed to

⁷"With that, you have correctly responded to the survey and accordingly coordinated with the disability support services."

⁸"Datenset" instead of "Datensatz", "Scala" instead of "Skala", "Modeswert" instead of "Modalwert", "Nominaldaten" instead of "Nominaldaten", "dealen" instead of "zu tun haben".

⁹"descriptive" instead of "deskriptive", "indeed" instead of "in der Tat"

be a help for them. Apart from disruptive moments just by the nature of encountering errors, the students could wonder if something else is meant and spend valuable time dissecting this hallucinated issue.

While Category 1 can cause specific issues, we would assume that the errors of Category 2 have a more concerning (long-term) impact. As long as the students confidently notice the errors, spending more time on the question and self-enforcing their knowledge could even strengthen their learning. On the contrary, when the feedback explained statistical terms and concepts, at least misleadingly or even plainly wrongly, this could negatively impact the students' understanding. Due to the formulative capacities of LLMs the explanations still *sound* correctly and students are probably not used to consider receiving erroneous information in the classroom. Often, these erroneous explanations were also not plain wrong but insufficient based on the context of the question and the overall learning goals. In these instances, the course instructor would just explain differently or add additional information.

4.3 Structure of the LLM generated feedback

Following the identification of errors in the feedback, we investigate in more detail how the feedback is composed. For doing so, we selected four questions that represent the range of tasks in this (mock) exam and coded the corresponding feedback based on our feedback categories derived from Hattie and Timperley (2007) and Ryan et al. (2020) (see subsection 2.3). As formulated in our research questions, we are especially interested in how the feedback does typically look like, how it does differ between tasks and students, and to which extent it aligns with educational feedback theories.

4.3.1 Overall Feedback Structure

Overall, across all 70 students and all four tasks, the most common feedback type is "response-oriented" (see fig. 2) which focuses on explaining why an answer is correct or incorrect, rather than merely stating its correctness ("right/wrong") or providing broader conceptual explanations ("conceptually-focused"). "Response-oriented" feedback occurs in 97 % of the 276 feedback instances that we coded manually – hence, in almost all feedback occurrences, bar a few exceptions. "Right/wrong" feedback, which simply indicates whether an answer is correct or not, is also very common (90 %), appearing in nearly nine out of ten feedback instances. "Self" feedback, which refers to personal evaluations or affirmations directed at the student (such as "Well done!", "Good job!" or "Great!") is present in more than two thirds of the feedback instances (68 %). "Process" (60 %) feedback, i.e., offering strategies and approaches for learning, still occurs in more than half of the instances. "Conceptually-focused" feedback, which provides additional explanations to deepen understanding, appears less frequently at 44 %. The - by far - least common feedback type is "self-regulatory" (21 %), which supports students in independently monitoring and correcting their own learning processes.

In addition to the frequency of feedback categories across all instances, we are also interested in the proportional volume of each category. Some types of feedback may occur in many instances, but only take up a small part of the output. Potentially, shorter feedback types might be perceived by students as less important and thus exert less impact on their learning. To investigate this aspect, we analyze the number of characters (including spaces) associated with each coded

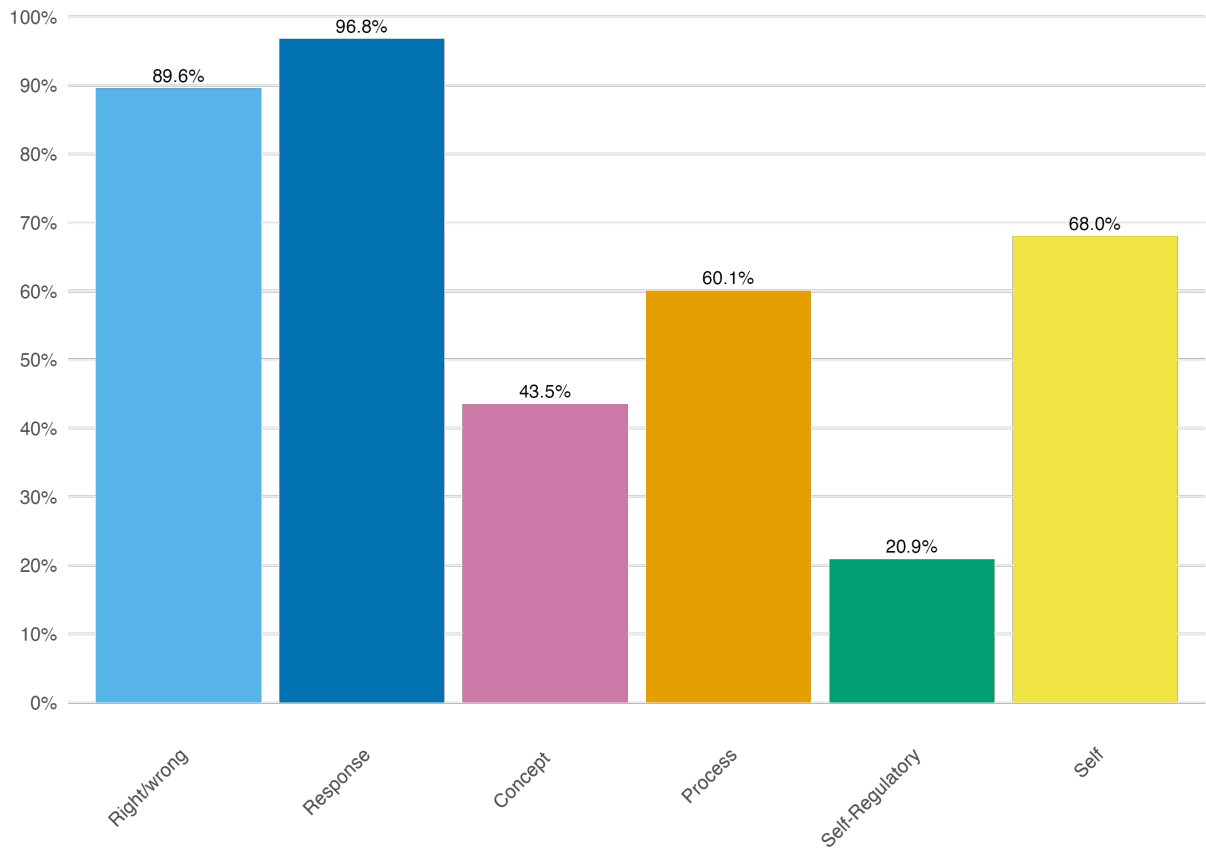


Figure 2: Frequency of feedback categories across all 70 students and 4 tasks.

feedback segment. "Response-oriented" is not just the most frequent feedback across the instances, it also accounts for the largest content share of all coded feedback. 43 % of the provided feedback belongs to this category (see fig. 3). "Conceptually-focused" feedback only comprises 21 % of the feedback coverage, "process" feedback 16 % and "right/wrong" feedback 13 %. "Self-regulatory" (4 %) and "self" (3 %) are the categories with the lowest feedback coverage. Interestingly, this suggests that merely considering the frequency of feedback categories is insufficient to fully understand how the feedback is composed and how it might be perceived by students. As expected, "right-wrong" is far more frequent than it is high in volume, as it takes not many words to state if an answer is correct or not. "Process" feedback only occurs in 60 % of the feedback instances, but takes the third-most overall space of all the feedback categories, indicating longer sequences of this type.

4.3.2 Feedback Structure by Task

We selected four tasks that capture the range of the mock exam and examined whether there are differences in the feedback structure between tasks:

- 1a: A multiple choice knowledge question about measures of central tendency.

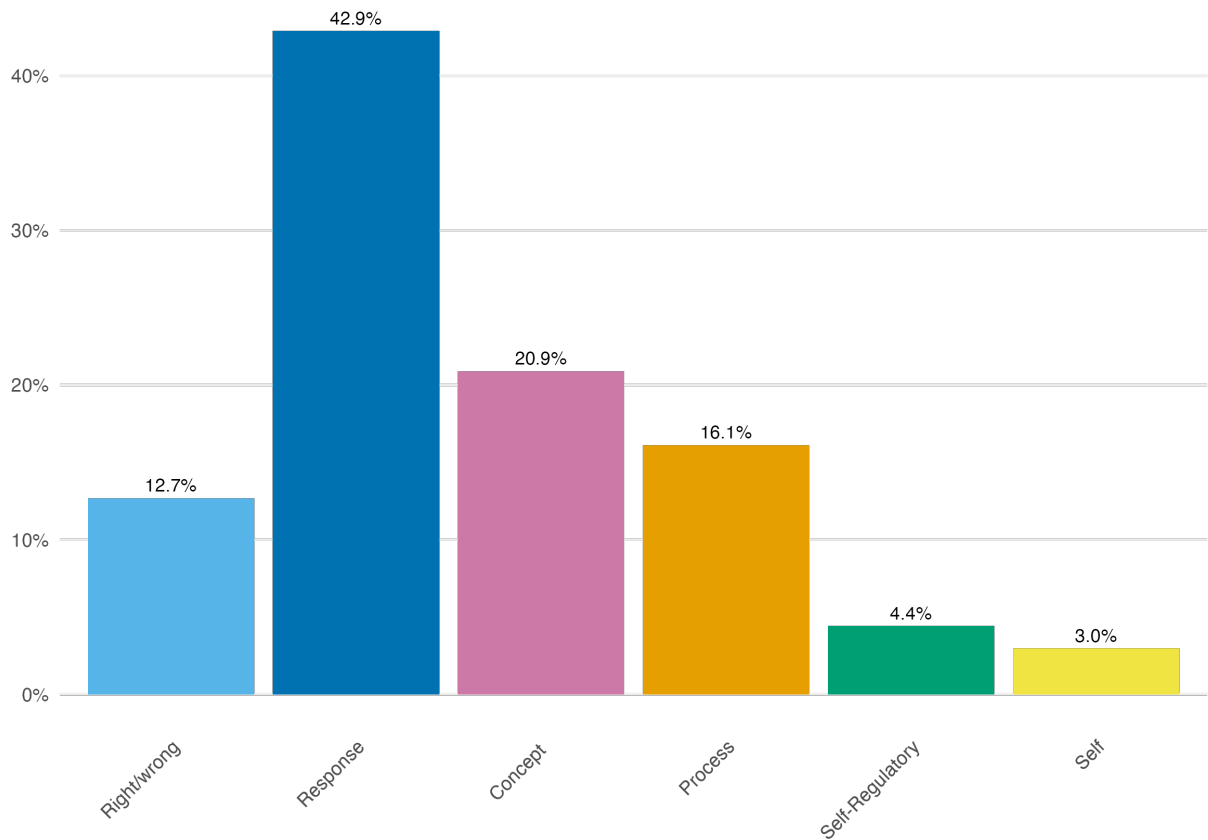


Figure 3: Coverage of feedback categories across all 70 students and 4 tasks.

- 2d: A graph interpretation question where students compare median and arithmetic mean.
- 3e: A table interpretation question where students have to decide between median and arithmetic mean based on outliers.
- 4a: A table calculation question where students have to calculate two sets of variances and standard deviation.

"Response-oriented" feedback is consistently the most frequent category across all 4 tasks and "right-wrong" feedback the second-highest (see fig. 4). Only in task 4 does "response-oriented" feedback fall under 95 %, which is also when "right-wrong" is the closest in frequency with only a difference of 1.5 percentage points.

Regarding the other feedback categories, we can observe stronger differences between tasks. Notably, "self-regulatory" feedback occurs in a very similar frequency of 15 - 17 % for three of the four tasks, but is much more common in one of the interpretation tasks (3e, 36 %). At the same time, "Conceptually-focused" feedback is considerably less frequent in this task (25 %) than in others (46 % - 58%). "Process" feedback, on the other hand, is notably more frequent for the calculation task (4a, 78 % vs. 49 - 59 %). Feedback of the "self" category is less prevalent for the interpretation tasks (2d: 61 %, 3e: 51 %) than for the knowledge (1a: 87 %) and calculation task (73 %).

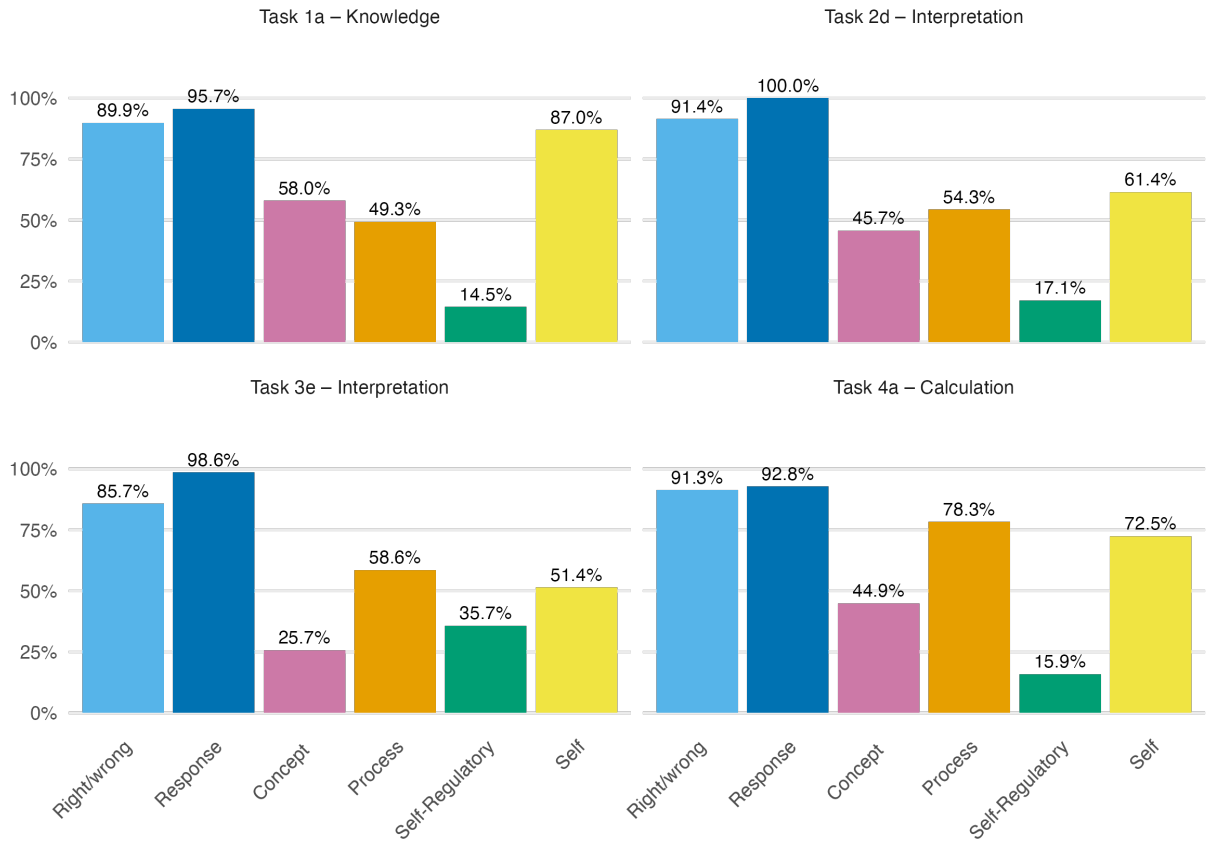


Figure 4: Frequency of feedback categories by task type.

Again, we supplement these findings on frequency by assessing the coverage of each category of feedback across the four tasks (see fig. 5). As above, the tasks show quite a similar coverage regarding the "right/wrong" category. However, there are bigger differences with response-focused feedback. The interpretation tasks 2d and 3e show a notably bigger coverage of "response-oriented" feedback. Whereas the coverage of process feedback is approximately equal across the tasks, we can see stronger variation with "conceptually-focused" and "self-regulatory" feedback: "conceptually-focused" feedback has a higher coverage for the knowledge and calculation task, while "self-regulatory" feedback is lower for those two compared to the interpretation tasks.

In other words: the knowledge task has a comparatively lower coverage of "response-oriented" and "process" feedback, but the highest for "conceptually-focused" and "self". The interpretation tasks have the highest coverage for "response-oriented" feedback, while the calculation task has the lowest coverage for "self-regulatory" feedback and low coverage for "response-oriented".

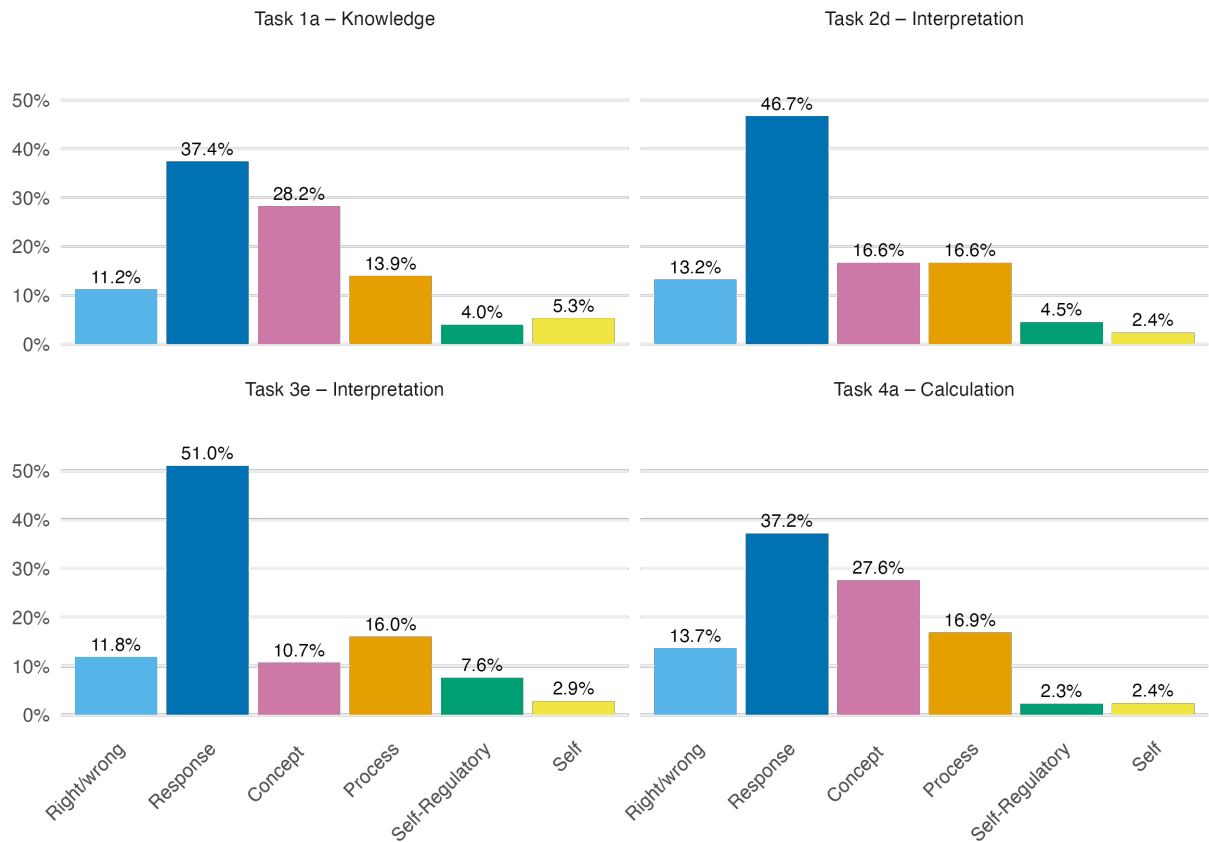


Figure 5: Coverage of feedback categories by task type.

4.3.3 Feedback structure by different points achieved

We are also interested in examining how the feedback differed between students. A key factor to differentiate students is the number of points they achieved on the respective task. The occurrence and coverage of different feedback types may vary depending on how well the student performed – similar to how a human instructor would provide different feedback to a student who answered incorrectly compared to one who gave a fully correct answer.

Figure 6 indicates that for "right/wrong" and "response-oriented" feedback, the occurrence does not vary meaningfully. However, the difference is very big for "conceptually-focused", "self-regulatory", and "self" feedback. Students, who achieved full points received considerably less often "conceptually-focused" and "self-regulatory" feedback and far more often "self" feedback; the opposite is true for students, who achieved zero points.

The same pattern is visible for the coverage of feedback based on the points achieved (see fig. 7: Students that perform better receive less "conceptually-focused" feedback. In contrast to the quite stable occurrence rates described above, "response-oriented" feedback shows a different pattern in coverage: Students achieving higher points receive more

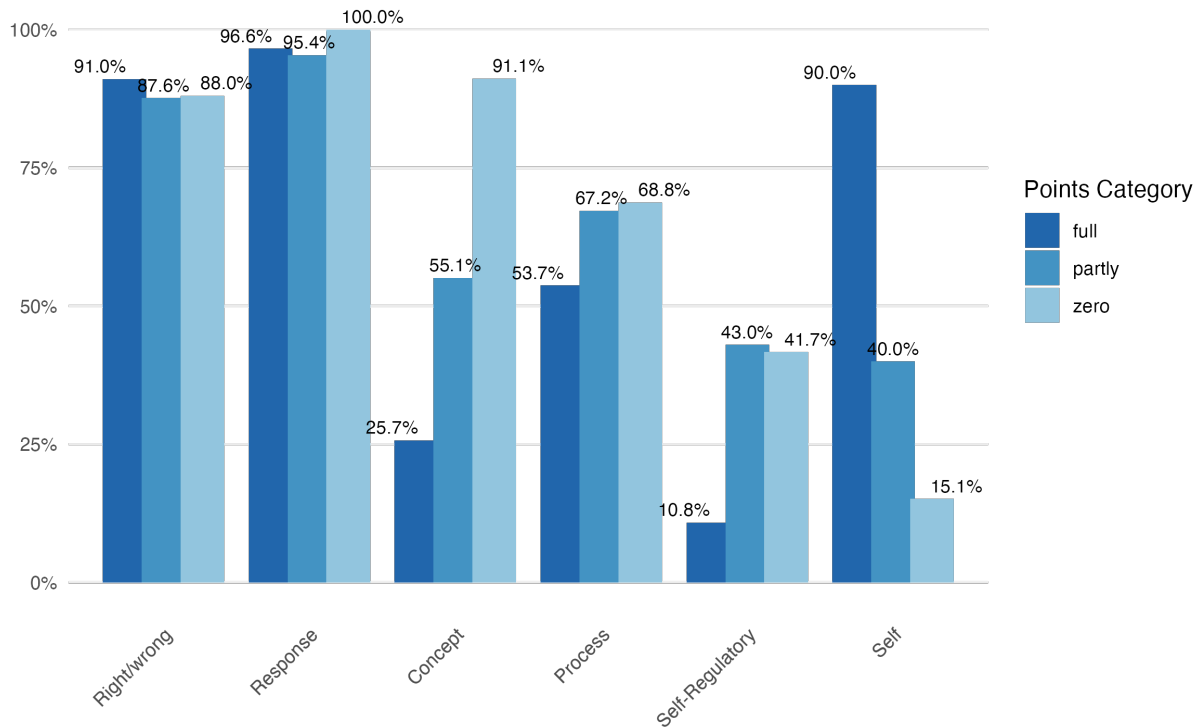


Figure 6: Frequency of feedback categories by points achieved.

"response-oriented" feedback than students performing worse. That seems counter-intuitive since one would assume that wrong answers typically need more clarification on why they are wrong.

4.3.4 Feedback structure by same points achieved

Another important aspect of examining the structure of LLM-generated feedback is the consistency of the feedback across different students who received the same points. Such differences may lead to unequal learning opportunities if students receive less information despite comparable performance. To analyze this, we qualitatively examine question 1a. As this is a multiple-choice question with pre-defined answer options, it affords the highest degree of comparability. In this task, students were supposed to select all measures of central tendency (mean, median, mode), excluding the fourth answer option, standard deviation. We will begin by examining the feedback provided to students who answered the question correctly, followed by an analysis of the feedback for those who made errors.

We observed substantial differences in feedback even among students who all answered this question correctly. As highlighted by the exemplary feedback in Table 3, the first obvious difference is the length of the different feedback instances. Examples 1 and 2 both contain the same feedback categories: "right/wrong", "response-oriented" and

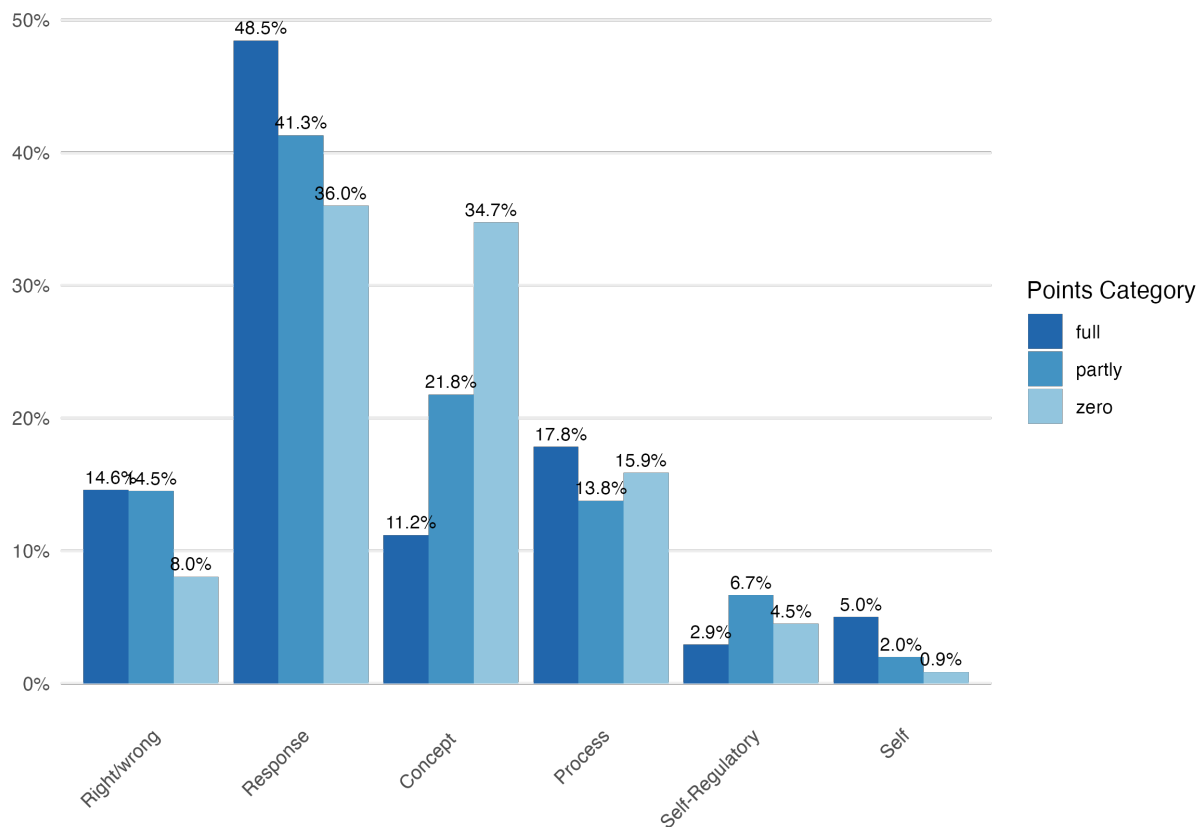


Figure 7: Coverage of feedback categories by points achieved.

"self" feedback. However, while Example 1 merely states that all selected answer options are correct because they are measures of central tendency, Example 2 adds another response-oriented piece of information: that the standard deviation, which was correctly omitted by the student, is a measure of dispersion. The student who received Example 1 did not get this information.

Example 3 shows an even starker contrast. While it includes the additional information on the incorrect answer option, similar to Example 2, it also adds "conceptually-focused" feedback. This example shows how naturally feedback can progress: First stating that the task was solved correctly ("right/wrong"), then specifying why the answer was correct or incorrect ("response-oriented"), and adding a short explanation of the concepts involved ("conceptually-focused"). While we might assume based on their correct answers that these students *know* which measures belong to which statistical concept, a brief reminder on their definition and what they are used for has the potential to further solidify their understanding.

#	Example feedback text
1	Very good, those are indeed all the measures of central tendency: the arithmetic mean (a), the median (b), and the mode (d). You solved this task correctly. Keep it up!
2	You solved the task correctly. The measures of central tendency in statistical analysis are indeed the arithmetic mean (a), the median (b), and the mode (d). The standard deviation (c), on the other hand, is a measure of dispersion. Keep it up!
3	You correctly identified the measures of central tendency. The arithmetic mean, the median, and the mode are indeed measures of central tendency. They give us an indication of where the "center" of a dataset is located. The standard deviation, which you correctly omitted, on the other hand, is a measure of dispersion, telling us how far the data are on average from the mean. Keep it up!

Table 3: English translations of selected feedback for task 1a (fully correct answers), highlighted by feedback category: "right/wrong" (light blue), "response-oriented" (darker blue), "conceptually-focused" (purple), "process" (red), "self-regulatory" (green), "self" (yellow).

Table 4 presents feedback responses for students who received partial points for task 1. In each case, students correctly identified “arithmetic mean” and “median” as measures of central tendency but omitted “mode.” This feedback is substantially longer than that for correct answers in Table 3, which is expected given the greater need for feedback when task-related knowledge gaps are determined. However, even among students with the same score, there are notable differences in the structure and content of feedback provided.

In Example 1, the student primarily receives “right/wrong” and “response-oriented” feedback, with the latter dominating. Two “self-regulatory” remarks are included, reminding the student to review all options carefully and to consider that multiple answers can be correct. By contrast, Examples 2 and 3 instead include brief elements of “process” feedback. In Example 3, this feedback assesses the student’s current learning status and encourages them to remember that “mode” is also a measure of central tendency. In Example 2, the feedback advises the student to revisit both concepts (measures of central tendency and dispersion) to deepen their understanding.

The biggest difference is the inclusion of “conceptually-focused” feedback, absent in Example 1 but prominent in Examples 2 and 3, matching the share of “response-oriented” feedback in Example 1. This shift is meaningful: instead of simply giving feedback on the answer options, the feedback provides explanations of the underlying statistical concepts.

Example 2 also identifies the incorrect answer (option 3) as a measure of dispersion and defines this term. In contrast, Example 3 omits this identification but offers definitions for all three measures of central tendency. This greater detail has the potential to highlight why these measures all belong to the same category, fostering a deeper understanding.

#	Example feedback text
1	Your answer is almost correct, but not complete. You correctly identified that the arithmetic mean (a) and the median (b) are measures of central tendency. But you missed the mode (d), which is also a measure of central tendency. The standard deviation (c) is a measure of dispersion, not of central tendency, which you correctly excluded. Always check all options carefully and remember that there can be multiple correct answers. Good work but don't forget that your answer is only complete when it addresses all parts of the question.
2	You already named two of the correct measures of central tendency – the arithmetic mean (a) and the median (b). However, you overlooked the mode (d), which is also a measure of central tendency. A small tip for the future: Measures of central tendency are statistical values that represent the central point of a distribution. These include not only the arithmetic mean and the median but also the mode. The standard deviation, on the other hand, is a measure of dispersion and indicates how much values deviate from the mean. Please review these concepts to develop a better understanding of the fundamental principles of statistics.
3	You forgot the mode (d) in your answer, which is also a measure of central tendency. It appears you have a basic understanding of statistical measures of central tendency, since you correctly identified the arithmetic mean and the median. Measures of central tendency are statistical indicators that give us an idea of where the “center” of a distribution lies. They include the arithmetic mean, the median, and the mode. Each of these provides a different definition of “center.” • The arithmetic mean is the average of the values. • The median is the middle value when all values are arranged in ascending order. • The mode is the value that occurs most frequently. Please make sure to include the mode as a measure of central tendency in your future answers. ¹⁰

Table 4: English translations of feedback for task 1a (partially correct answers), highlighted by feedback category: "right/wrong" (light blue), "response-oriented" (darker blue), "conceptually-focused" (purple), "process" (red), "self-regulatory" (green), "self" (yellow).

5 Discussion and Conclusion

Our results provide an in-depth analysis of LLM-generated feedback in a real (statistics) classroom setting with 70 students working on a mock exam. This specific course setting usually does not allow for individual feedback for every student. However, any errors or misleading responses could significantly hinder students' learning progress. Subsequently, our approach was to use the text-generating capabilities of LLMs, while supplying the correct answers from the human instructors as contextual input for the LLMs to mitigate errors. To investigate if we can trust LLMs to deliver feedback of high quality in this context, we guided our study with two principal research questions:

1. To what extent does the LLM produce errors while giving feedback when the instructors supply the correct answers as contextual input?
2. What is the structure of the feedback generated, including its variations across tasks and students, and its alignment with established educational feedback theories?

¹⁰This "process" categorization is an example for the fringe cases that are expected with basing the categories on Hattie and Timperley (2007) where the levels are not completely disjunct but rather interrelated. In this case, the reminder to include it in the future answers can be interpreted as "self-regulatory", but as it rather hints towards reviewing the mode as a measure of central tendency and, hence, suggesting how to approach this statistical measure, it was labeled as "process" level. This can be compared with the rather different "self-regulatory" feedback in the first example of the table.

For our first research question, we found that 6.99% (n=167) of the 2,389 feedback instances contained errors. Given the sensitivity of the setting, we applied a broad definition of errors, encompassing any occurrence potentially confusing, distracting, or misleading for students. While a considerable number of these errors are simple one-word mistakes or erroneous translations, some include inaccurate or misleading explanations of statistical concepts. There were some tasks of the mock exam where errors occurred more frequently, but we found errors in the majority of the 38 task. Only 13 tasks contained no errors at all. We differentiated the most common errors into two categories based on their potential effect on students. Category 1 includes technical mistakes that are easily identifiable by students, such as a wrong handling of currencies (not accepting 0.30 Euro as a response for the correct answer 30 cents). While they still have the potential to be slightly disruptive and distracting, their negative impact is likely more limited as students can readily recognize them. On the other hand, Category 2 comprises more subtle issues, such as conceptually misleading explanations. These have a greater risk of introducing misconceptions without the student noticing, resulting in a 'silent' negative effect on learning.

To address our second research question on feedback structure, we conducted a detailed content analysis of four exam tasks. Our analysis scheme is based on the established Hattie and Timperley (2007) framework, which we expanded by detailing the task-level feedback into three more subcategories drawn from Ryan et al. (2020). We examined the structure of feedback both by frequency of occurrence and by text volume. Overall, we found expected outcomes such as "right/wrong" and "response-oriented" feedback being the most frequent across all students and tasks. Meaning, that in the vast majority of cases, the students got told if their answer is (in)correct ("right/wrong") and why ("response-oriented"). "Conceptually-focused" feedback, which explains the underlying concepts of a task, occurred less frequently but tended to be more elaborate when present. As these three categories are, in our approach, a dissection of the "task" level of Hattie and Timperley (2007), it aligns strongly with previous empiric research that most feedback addresses the task level both for human and LLM contexts (Dai et al. (2023)). Our subdivision in three different types of the task level as the basis for a quantified analysis of qualitative data demonstrates that there can be more nuance to this feedback level. Wisniewski et al. (2020) showed that the impact of feedback is substantially influenced by the conveyed information content. This information content should rise strongly in our sub-categories from "right/wrong", to "response-oriented", and finally to "conceptually-focused".

"Process" feedback appeared in over half of the feedback instances, although often just as brief phrases reflecting the student's current understanding and less often suggesting actual learning strategies. "Self-regulatory" feedback was by far the least frequent category, indicating students rarely received guidance on how to assess or monitor their own learning. This rare occurrence is consistent with previous quantifications of feedback levels (Dai et al., 2023) and can mark a flaw in feedback design as this "self-regulatory" level can be considered as one of the most useful feedback levels (Wisniewski et al. (2020)).

"Self" feedback, such as praise or personal comments (described as rarely effective by Hattie and Timperley (2007)) was the third most common category. However, it usually consisted of very brief remarks of praise ("Good job!", "Great!"), resulting in the lowest overall text volume among all categories. The prevalence of such "self" feedback

flattery aligns with evidence that LLMs can tend to over-optimize for human approval or validation at the expense of correctness or faithfulness, known under the term *sycophancy* (Sharma et al., 2025). This is amplified by cases in our feedback data where the LLM keeps flattering even when the student is not giving correct answers ("*Good work, but don't forget that your answer is only complete when it addresses all parts of the question.*") While this can be interpreted as encouragement, it could also be distracting from the actual suggestions that are given. Hattie and Timperley (2007, p. 96) argued that praise "is unlikely to be effective, because it (...) too often deflects attention from the task." LLM studies indicate that *sycophancy* is potentially a by-product of human-preference training, meaning that humans and, consequentially, preference models favor affirmative responses (Sharma et al., 2025). This ambivalence is yet another connection to feedback research, where "self" feedback remains a complex and multi-layered topic. While in a recent revisit of the Hattie and Timperley (2007) model, "self" feedback was rated least useful by students (Mandouit and Hattie, 2023), other learner-centered research argues that "students still want to receive praise to facilitate their motivation moving forward" (Van Boekel et al., 2023, p. 3399).

We further analyzed feedback by the type of task (knowledge, interpretation, calculation), but found little interpretable patterns across these types. For example, one might expect interpretation tasks to elicit more "conceptually-focused" feedback, as they require deeper engagement with statistical concepts. Yet, one of the two interpretation tasks had the lowest frequency of this feedback type among the four analyzed.

Clearer patterns could be observed in the relation between feedback structure and the number of points achieved. Students receiving partial or no points received considerably more frequently "conceptually-focused" and "self-regulatory" feedback and less frequently "self" feedback. "Right/wrong" and "response-oriented" feedback remained relatively constant regardless of the performance. This is intuitive: students who answered correctly still received validation and praise, but less elaboration and guidance compared to those who struggled.

We also explored how consistent feedback was for students with similar performance. To do this, we qualitatively analyzed feedback for task 1a and found substantial variation, both among students who answered correctly and those who did not receive full points. In both groups, a key difference was whether or not students received "conceptually-focused" feedback. While this may be more negligible for students who answered correctly, it is potentially more problematic that students with identical incorrect answers sometimes received detailed explanations, while others did not.

Additionally, our qualitative analysis showed that even within the same category, feedback could differ strongly in content. For example, one feedback instance provided "conceptually-focused" explanations for both statistical concepts involved (central tendency and dispersion), while another addressed only central tendency, but elaborated in detail on the specific measures (mean, median, and mode). This indicates that even when the feedback structure is consistent, the actual content can vary considerably.

Evaluating this feedback in depth presents a challenge, especially given its notable variance. The feedback levels of Hattie and Timperley (2007) are deliberately overlapping, as the most effective feedback should progress fluidly across

levels. This requires judgment calls from coders, as some text segments may align with different levels depending on the perspective. Overall, our analysis shows that the levels, in combination with our extension of the task level based on the Ryan et al. (2020) categories, provide a suitable framework for accurately describing the structure and composition of feedback.

Moreover, we believe this approach can serve as a foundation and guide for future feedback design. While many studies evaluate feedback (often drawing on Hattie and Timperley (2007)), they tend to focus on more elaborate tasks such as creative writing or project reports. These task types naturally provide more content and context for feedback than the compact, individual items used in our exam. Nonetheless, this particular context is critically important: in large-scale courses, students often receive no individual feedback at all. In this light, we must carefully weigh whether the presence of occasional errors or structural inconsistencies in LLM-generated feedback outweighs the alternative of providing no feedback whatsoever. One could also question how consistent instructor-generated feedback would actually be, and whether it would always avoid ambiguity or misinterpretation. Many instructors are not familiar with feedback theory and may not consistently provide the most effective and pedagogically sound comments. Still, it remains essential to hold automated systems to a higher standard, when the responsible human is no longer part of the feedback loop in order to intervene in case of problems.

Providing feedback via LLMs opens the door to offering individualized feedback where none would typically be available. In addition, it enables a shift from reactive, ad-hoc responses to a more deliberate, design-oriented approach to feedback. One can assume that instructors under common time constraints rarely design individualized feedback systematically and apply it consistently across all students. Instead, they might tend to comment on responses based on their teaching experience and personal didactic values. By leveraging LLMs, we gain the opportunity to invert this process: first defining what effective feedback should look like, and then deploying it consistently based on these predefined principles and frameworks. The framework presented in this paper can serve precisely as a foundation for such design. The key lies in making these feedback decisions consciously. For example, it may be appropriate to provide more detailed explanations to students who answered incorrectly than to those who answered correctly. But such decisions should be intentional, not arbitrary or inconsistent across similarly performing students. This engagement with deliberate feedback design also provides opportunities for instructors to sharpen exercise and exam design, ensuring that each task genuinely assesses the intended learning outcomes.

In this pilot study, we used the platform in an out-of-the-box configuration, with a generic prompt not specifically aligned with educational feedback theories. This setup reflects a realistic scenario in which instructors adopt such platforms with minimal customization. A promising avenue for future research is to evaluate how aligning prompts, incorporating retrieval-augmented generation (RAG), or applying fine-tuning with high-quality human feedback and our empirical findings affects the effectiveness of feedback. While RAG extends large language models by dynamically retrieving and integrating relevant external information into the generation process, fine-tuning adapts the model's parameters to domain-specific data, thereby improving accuracy and contextual relevance. Exploring these approaches

in combination with high-quality human feedback could provide a more robust foundation for generating effective and pedagogically sound feedback.

Recent feedback research suggests to more strongly emphasize feedback literacy, meaning the learner's perception of feedback (Carless and Boud, 2018; Mandouit and Hattie, 2023; Weidlich et al., 2025). We reflected this by conducting an evaluation survey: students responded in general very favorably to the implementation of the LLM-generated feedback. Particularly, 93 % rated the feedback rather or very useful. Further, our survey data indicates that most students already use ChatGPT or other LLMs on their own for studying. Introducing it as part of an educational tool in the classroom with the opportunity to answer questions and address more common issues, provides an additional security net compared to students using those tools autonomously without supervision. In our error analysis, we mainly took the interpretation from the students' side, contemplating the weight of the errors by how easily students could detect them. Future work could deepen this literacy focus by embedding a feedback loop in which students review each feedback response, flagging potential errors and rating the helpfulness of the feedback. This would also align with the argument that feedback should be an ongoing dialogue (Mandouit and Hattie, 2023). The *StudyLabs* platform offered students the opportunity to keep chatting with the feedback they received for every single task. This feature was very rarely used by the students, but could be a potentially valuable addition if used more consequentially and intentionally.

The present paper is not only a theory-informed technical and qualitative evaluation of AI-generated feedback in an authentic classroom setting. The feedback evaluation framework that we extended from Hattie and Timperley (2007) in combination with Ryan et al. (2020) is not limited to the assessment of LLM feedback, it is equally applicable to human-generated feedback. Through this combined framework, we conducted a fine-grained analysis that dissects common task-level feedback in greater detail and offers new insight into feedback composition. This demonstrates how engaging with LLM-generated feedback goes beyond merely evaluating the implementation of new AI systems, it opens up broader discussions about the nature, quality, and design of educational feedback at its core.

Declarations

Availability of data and materials

We published the anonymized pre- and post-survey data, analysis code and (translated) codebooks: [Link to OSF repository](#). The repository is currently published anonymously for blinded review. Please note that we published the GPT-usage and StudyLabs-experience data from the survey (as reported in Results Section 4.1), while not publishing sociodemographic information, study program data, and open-text responses to prevent potential re-identification through variable combinations.

We attached a translation of the mock exam including task type categorizations as part of this submission as supplementary material.

Regarding the potential publication of the feedback data from the StudyLabs platform (including scoring and feedback) and our subsequent categorization data, we are still in contact with ZAVI as the data owners. We will update this statement when a decision is reached.

Funding

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

Generative AI and AI-assisted technologies in the writing process

Generative AI (genAI) in the form of Grammarly and ChatGPT was used for language improvements. Further, ChatGPT was used for assistance in translation, literature discovery, R programming, and LaTeX coding. Generative AI tools were **not** used at any stage to process raw data or draft the manuscript.

Acknowledgments

We thank ZAVI for providing the software as well as the data.

References

- Ayanwale, M. A., Adelana, O. P., and Odufuwa, T. T. (2024). Exploring STEAM teachers' trust in AI-based educational technologies: a structural equation modelling approach. *Discover Education*, 3(1):44.
- Carless, D. and Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8):1315–1325. Publisher: SRHE Website _eprint: <https://doi.org/10.1080/02602938.2018.1463354>.
- Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.-S., Gasevic, D., and Chen, G. (2023). Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT.
- Er, E., Akçapınar, G., Bayazit, A., Noroozi, O., and Banihashem, S. K. (2025). Assessing student perceptions and use of instructor versus AI-generated feedback. *British Journal of Educational Technology*, 56(3):1074–1091. _eprint: <https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13558>.
- Estevez-Ayres, I., Callejo, P., Hombrados-Herrera, M. A., Alario-Hoyos, C., and Delgado Kloos, C. (2024). Evaluation of LLM Tools for Feedback Generation in a Course on Concurrent Programming. *International Journal of Artificial Intelligence in Education*.
- Feldman-Maggor, Y., Cukurova, M., Kent, C., and Alexandron, G. (2025). The Impact of Explainable AI on Teachers' Trust and Acceptance of AI EdTech Recommendations: The Power of Domain-specific Explanations. *International Journal of Artificial Intelligence in Education*.
- Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., and Drachsler, H. (2024). From the Automated Assessment of Student Essay Content to Highly Informative Feedback: a Case Study. *International Journal of Artificial Intelligence in Education*, 34(4):1378–1416. FE + FT Example.
- Hattie, J. and Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

- Hellas, A., Leinonen, J., Sarsa, S., Koutcheme, C., Kujanpää, L., and Sorva, J. (2023). Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, pages 93–105, Chicago IL USA. ACM. FE: LLM Evaluation.
- Jacobsen, L. J. and Weber, K. E. (2023). The Promises and Pitfalls of LLMs as Feedback Providers: A Study of Prompt Engineering and the Quality of AI-Driven Feedback. FE: LLM Evaluation.
- Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., and Meyer, J. (2024). Empirische Arbeit: Comparing Generative AI and Expert Feedback to Students' Writing: Insights from Student Teachers. *Psychologie in Erziehung und Unterricht*, 71(2). FE: LLM Evaluation.
- Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., and Gehringer, E. (2024). On Assessing the Faithfulness of LLM-generated Feedback on Student Assignments. Publisher: International Educational Data Mining Society.
- Jürgensmeier, L. and Skiera, B. (2024). Generative AI for scalable feedback to multimodal exercises. *International Journal of Research in Marketing*, 41(3):468–488.
- Li, T. W., Hsu, S., Fowler, M., Zhang, Z., Zilles, C., and Karahalios, K. (2023). Am I Wrong, or Is the Autograder Wrong? Effects of AI Grading Mistakes on Learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, pages 159–176, Chicago IL USA. ACM.
- Liang, X., Song, S., Zheng, Z., Wang, H., Yu, Q., Li, X., Li, R.-H., Xiong, F., and Li, Z. (2024). Internal Consistency and Self-Feedback in Large Language Models: A Survey. arXiv:2407.14507 [cs] version: 1.
- Lipnevich, A. A. and Panadero, E. (2021). A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. *Frontiers in Education*, 6. Publisher: Frontiers.
- Liu, J., Jiang, B., and Wei, Y. (2025). LLMs as Promising Personalized Teaching Assistants: How Do They Ease Teaching Work? *ECNU Review of Education*, 8(2):343–348.
- Lohr, D., Keuning, H., and Kiesler, N. (2025). You're (Not) My Type- Can LLMs Generate Feedback of Specific Types for Introductory Programming Tasks? *Journal of Computer Assisted Learning*, 41(1):e13107. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.13107>.
- Lyu, W., Zhang, S., Chung, T., Sun, Y., and Zhang, Y. (2025). Understanding the practices, perceptions, and (dis)trust of generative AI among instructors: A mixed-methods study in the U.S. higher education. *Computers and Education: Artificial Intelligence*, 8:100383.
- Makransky, G., Shiwalia, B. M., Herlau, T., and Blurton, S. (2025). Beyond the "Wow" factor: Using Generative AI for Increasing Generative Sense-Making. *Educational Psychology Review*, 37(60). FE: LLM Evaluation.
- Mandouit, L. and Hattie, J. (2023). Revisiting “*The Power of Feedback*” from the perspective of the learner. *Learning and Instruction*, 84:101718.
- Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution.

- Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., and Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.
- Nazaretsky, T., Cukurova, M., and Alexandron, G. (2022). An Instrument for Measuring Teachers' Trust in AI-Based Educational Technology. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, pages 56–66, New York, NY, USA. Association for Computing Machinery.
- Phung, T., Pădurean, V.-A., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., Singla, A., and Soares, G. (2023). Generative AI for Programming Education: Benchmarking ChatGPT, GPT-4, and Human Tutors. arXiv:2306.17156 [cs].
- Phung, T., Pădurean, V.-A., Singh, A., Brooks, C., Cambronero, J., Gulwani, S., Singla, A., and Soares, G. (2024). Automating Human Tutor-Style Programming Feedback: Leveraging GPT-4 Tutor Model for Hint Generation and GPT-3.5 Student Model for Hint Validation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 12–23, Kyoto Japan. ACM.
- Roest, L., Keuning, H., and Jeuring, J. (2023). Next-Step Hint Generation for Introductory Programming Using Large Language Models. arXiv:2312.10055 [cs].
- Ryan, A., Judd, T., Swanson, D., Larsen, D. P., Elliott, S., Tzanetos, K., and Kulasegaram, K. (2020). Beyond right or wrong: More effective feedback for formative multiple-choice tests. *Perspectives on Medical Education*, 9(5):307–313.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. (2025). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548 [cs].
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., and Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91:101894.
- Steiss, J., Tate, T. P., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., and Uci, M. W. (2023). Comparing the Quality of Human and ChatGPT Feedback on Students' Writing. FE: LLM Evaluation.
- Sun, D. L., Harris, N., Walther, G., and Baiocchi, M. (2014). Peer assessment enhances student learning. arXiv:1410.3853 [stat].
- Topali, P., Cobos, R., Agirre-Uribarren, U., Martínez-Monés, A., and Villagrà-Sobrino, S. (2024). 'Instructor in action': Co-design and evaluation of human-centred LA-informed feedback in MOOCs. *Journal of Computer Assisted Learning*, 40(6):3149–3166. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.13057>.
- Van Boekel, M., Hufnagle, A. S., Weisen, S., and Troy, A. (2023). The feedback I want versus the feedback I need: Investigating students' perceptions of feedback. *Psychology in the Schools*, 60(9):3389–3402. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/pits.22928>.

- Venter, J., Coetzee, S. A., and Schmulian, A. (2025). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment & Evaluation in Higher Education*, 50(4):516–536. Publisher: SRHE Website
_eprint: <https://doi.org/10.1080/02602938.2024.2415649>.
- Viberg, O., Cukurova, M., Feldman-Maggor, Y., Alexandron, G., Shirai, S., Kanemune, S., Wasson, B., Tømte, C., Spikol, D., Milrad, M., Coelho, R., and Kizilcec, R. F. (2024). What Explains Teachers' Trust in AI in Education Across Six Countries? *International Journal of Artificial Intelligence in Education*.
- Wan, T. and Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*, 20(1):010152. FE + PE: Evaluation.
- Weidlich, J., Fink, A., Frey, A., Jivet, I., Gombert, S., Menzel, L., Giorgashvili, T., Yau, J., and Drachsler, H. (2025). Highly informative feedback using learning analytics: how feedback literacy moderates student perceptions of feedback. *International Journal of Educational Technology in Higher Education*, 22(1):43.
- Wisniewski, B., Zierer, K., and Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, 10. Publisher: Frontiers.

A Appendix

A.1 Most common feedback errors

Table 5: Most common feedback errors with example text (translated)

Task	#	Error	Example (translated from German)
7a	10	Does not accept “0.3021 Euro” for “30 cent”	“You correctly recognized the concept of the estimated coefficient in this context, but in this case there seems to have been a small mix-up. The interpretation of the coefficient 0.3021 actually means an increase of 0.3021 Euro (about 30 cents) per one-unit increase in the predictor, not “30 cent” as a separate unit. Clarifying this helps ensure a correct understanding of the matter. (...)”
3f	34	Ignores new table for the exercise, refers to the previous exercise description instead	“Although your calculation is basically correct, you apparently used the wrong table. In the exercise prompt, the values 23, 41, 20, and 13 are given. (...)”
6b	14	Shows correct formula for calculating χ^2 from the provided data but explains a different formula (first computing expected values)	“(...) The χ^2 calculation consists of several steps that sum (observed – expected)² divided by the expected frequency across all cells. In this case, with the given data, the calculation would be: $\chi^2 = 3467 \cdot \left(\frac{191^2}{397 \cdot 1760} + \dots + \frac{1120^2}{2374 \cdot 1707} - 1 \right)$ This yields a value of 13.59 for the χ^2 statistic. (...)”
1e	6	Missing distinction between Spearman and Pearson correlations (creates the impression that Spearman can only be used for metric variables)	“(...) The Pearson correlation (answer b) and the Spearman correlation (answer a), on the other hand, measure the relationship between two continuous variables and are therefore not suitable in this case.”
3c, 3d	8	Misleading confusion between ordinal and nominal variables	“(...) Regardless of whether the data were collected on a nominal or an ordinal scale, the median can be estimated as long as the data can be ordered. (...)”

Task	#	Error	Example (translated from German)
7h	30	Assumes the wrong scale of the independent variable (variable describes the dichotomous existence of children, not their number)	“(...) 1) Identify the independent variables. In this case they are income and the number of children in the household. 2) Identify the dependent variable. This is the amount a household spends on a summer vacation. 3) Use the given model for prediction. If you plug in the values, you obtain the predicted expenses for the vacation. Now take the given values and plug them into the formula: income = 2700 Euro and children = 2. This yields the following calculation: $259.02195 + 0.27891 \cdot 2700 - 66.71902 \cdot 2 = 945.3599.$ (...)”

A.2 Abbreviations

AI	Artificial Intelligence
e.g.	exempli gratia
i.e.	id est
GPT	Generative Pre-trained Transformer
LLM	Large Language Model