Feedback on Feedback: Student's Perceptions for Feedback from Teachers and Few-Shot LLMs

Sylvio Rüdian
Department of Computer Science Education
Humboldt-Universität zu Berlin
Berlin, Berlin, Germany
ruediasy@informatik.hu-berlin.de

Julia Podelo
Professional School of Education
Humboldt-Universität zu Berlin
Berlin, Berlin, Germany
julia.podelo@hu-berlin.de

Jakub Kužílek
Department of Computer Science Education
Humboldt-Universität zu Berlin
Berlin, Berlin, Germany
jakub.kuzilek@hu-berlin.de

Niels Pinkwart
Educational Technology Lab
German Research Center for Artificial Intelligence (DFKI)
Berlin, Berlin, Germany
niels.pinkwart@hu-berlin.de

Abstract

Large language models (LLMs) can be a valuable resource for generating texts and performing various instruction-based tasks. In this paper, we explored the use of LLMs, particularly for generating feedback for students in higher education. More precisely, we conducted an experiment to examine students' perceptions regarding LLM-generated feedback. This has the overall aim of assisting teachers in the feedback creation process. First, we examine the different student perceptions regarding the feedback that students got without being aware of whether it was created by their teacher or an LLM. Our results reveal that the feedback source has not impacted how it was perceived by the students, except in cases where repetitive content has been generated, which is a known limitation of LLMs. Second, students have been asked to identify whether the feedback comes from an LLM or the teacher. The results demonstrate, that students were unable to identify the feedback source. A small subset of indicators has been identified, that clearly revealed from whom the feedback comes from. Third, student perceptions are analyzed while knowing that feedback has been auto-generated. This examination indicates that generated feedback is likely to be met with resistance. It contradicts the findings of the first examination. This emphasizes the need of a teacher-in-the-loop approach when employing auto-generated feedback in higher education.

Keywords

Large Language Models, Prompt Engineering, Feedback Indicators, Language Learning

ACM Reference Format:
Sylvio Rüdian, Julia Podelo, Jakub Kužílek, and Niels Pinkwart. 2025. Feedback on Feedback: Student's Perceptions for Feedback from Teachers and Few-Shot LLMs. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706479

LAK 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706479

1 Introduction

Learning represents the primary goal of any educational endeavor. Without delving into the numerous attempts to define the term "learning," it can certainly be stated in a general sense that learning involves acquiring new abilities, skills, and knowledge and occurs in various life contexts. So, feedback can be seen as "a 'consequence' of performance"[21] and therefore as a fundamental element of any learning process: For instance, young children touch a hot stove and receive biological feedback in the form of pain, which usually leads quite quickly to a learning outcome: The stove is hot, and hot hurts. Nevertheless, the concept of feedback has surprisingly only been present in educational discourses for some decades and has thus only recently become a focus of empirical educational research [11]. In this regard, feedback, especially in verbal form, proves to be extremely effective: Verbal feedback enhances learners' performance more than grades or graded comments alone [29]. While grades suggest the end of the learning process, feedback emphasizes the ongoing learning process [19]. In this context, feedback is understood as "a process through which learners make sense of information from various sources and use it to enhance their work or learning strategies" [4]. According to Hattie [19, 20], feedback is an essential element that can improve learning and teaching. It can communicate the gap between the learning goals leading teachers expectations, and learner skills. This can be referred to corrective feedback, where especially mistakes or misconceptions are communicated [30]. In general, feedback can be related to the task, or process level, to learners self-regulation, and one's self [66]. Feedback can positively impact the student achievement [21].

However, for this to be effective, the recipient of the feedback needs a certain level of feedback literacy. According to Carless and Boud, learners with good feedback literacy are able to appreciate feedback processes and their role within them, make their own judgments, and process the emotional consequences of feedback, all with the central goal of translating received feedback into concrete actions [4]. To develop this form of feedback literacy, learners require numerous active and passive feedback situations, as well as educators who are trained in the forms, functions, and impacts

of feedback. Given the proven effects of feedback on learning outcomes [19], this underscores the necessity of increasingly focusing on feedback literacy in teacher education for future educators.

At the same time, such an endeavor also places a certain time burden on teaching staff: Providing constructive feedback that promotes learning and offering it in a sustainable manner often requires more time than simply assigning a grade [3]. The efficacy of feedback generated under temporal constraints in fostering optimal learning outcomes remains a topic of debate.

Observing the recent technology, large language models (LLMs) have been arising, a promising approach that could be utilized to assist the process of creating feedback. They can handle textual learner submissions, for which feedback can principally be generated based on an instruction [65]. Such automatically generated feedback is the scope of this paper, focusing on the students' perspectives, especially their perceptions and feedback literacy.

2 Related Work

The number of studies utilizing LLMs in different scenarios is overwhelming [16]. They can be found in medicine [39], education [25], finance [31], engineering [24], and others. Subsequently, LLMs are of interest in industry, and academia [68] due to its capabilities. Especially LLMs assist in writing tasks, provide editing suggestions, e.g. in proofreading, can summarize or translate texts, extract key points, identify unexplored aspects, but they can also write documents, help with problem solving, and more [25].

Access to an LLM involves using so-called prompts [65]. Such prompts are LLM-instructions with context to conduct a (textual) task. The process of identifying prompts that generate desired outputs is called prompt engineering [35], or prompting [34]. Researchers identified a bunch of prompts, that can be employed for different scenarios, and different domains [65]. Nevertheless, many tasks should be conducted cautiously due to the lack of interpretability and the potential for semantically faulty LLM outputs [15]. For example, in a study to generate tasks for language learners, an LLM has been instructed to create a text with a given set of semantically connected words [48]. The evaluation of the results revealed that an LLM often creates correct outputs, but sometimes, generated texts suddenly switch the context, making them unusual, so that they still need revision. Hence, there is still a need for a human-in-the-loop. Furthermore, LLM outputs can be biased [41]. In general, the development of effective prompts in research relies on an iterative process of exploratory prompt engineering, wherein optimal formulations are discovered through a heuristic "trial and error" approach. [50]. Then, there must be an evaluation process to identify whether the LLM component can successfully be integrated into a particular scenario.

In research, an LLM can be evaluated based on datasets, containing pre-defined inputs, and outputs. To name some of them: AGIEval [69], AlpacaFarm [12], BBQ [46], BoolQ [8], Codex HumanEval [7], GSM8K [9], HELM [32], MMLU [22], MultiPL-E [5],ToxiGen [18], WinoGrande [51], and others. They are employed to identify how well an LLM performs in generating known outputs and serve as benchmarks that allow LLM comparison [6]. However, if a set of prompts is designed that conducts a specific task, aiming to be integrated into a concrete scenario, then an appropriate evaluation metric must be identified to make a decision [50].

In this study, an LLM is employed to generate feedback in the context of language education.

Prior to the advent of LLMs, researchers concentrated on developing Automatic Writing Evaluation (AWE) systems, designed to furnish constructive feedback for open-writing assignments [57]. The implementation of such tools has been shown to significantly influence student outcomes, particularly in language teaching [45]. Nevertheless, their application within traditional classroom settings has thus far been limited to the provision of corrective feedback [23], underscoring a need for more comprehensive integration. Furthermore, a notable discrepancy exists between the erroneous content detected by these tools and that identified by teachers, with inter-rater reliability being a significant concern [47]. Despite these limitations, the capabilities of LLMs remain auspicious, warranting further exploration and development to fully harness their pedagogical potential.

To successfully integrate a feedback generator into higher education settings, its usefulness, and acceptance must be identified. The evaluation criteria to determine whether feedback is appropriate, leads to the students' expectations [40]. There are different indicators, on which feedback quality can be measured. According to Nicol and Macfarlane-Dick [42], seven principles can be found that support self-regulation in students: 1) defining good performance, 2) facilitates self-reflection, 3) information about their learning, 4) encourage dialogue, 5) encourage positive motivation, 6) actions to close knowledge gaps, 7) improve teaching. Subsequently, it is not surprising that feedback should be correct and understandable so that students can benefit from it. Depending on the purpose, and time point when feedback is provided to learners, we distinguish between formative, and summative feedback [13]. Formative feedback is provided within the learning process [54], while summative feedback is mainly provided at the end of the learning process, e.g. for grading [26]. Next to the time when feedback is provided, it can be distributed via different channels. The study of Mulliner and Tucker [40] revealed, that individual face-to-face feedback is most effective for students, followed by individual written feedback, while typed, or hand-written feedback has a similar effect for students. Also, typed feedback is the most student-preferred option. For teachers, typed feedback is more efficient than creating handwritten feedback [40]. Subsequently, focusing on typed feedback is a good trade-off between effectiveness, and students' preferences.

According to Van der Kleij et al. [63], the effect of elaborated feedback is higher than for the corrective one. Whether only one type of mistakes should be included in feedback, or the entire bunch of mistakes, is not determined, and may vary on the domain [53]. Pankiewicz et al. [44] identified, that LLM-generated formative feedback "did not significantly impact students' affective state". Wei et al. [10] measured the agreement between LLM-generated feedback, and a teacher. The LLM was capable of generating more detailed feedback, it was more accurate when summarizing the student's performance, and its overlap to teacher feedback has been high. However, their evaluation focused only on readability metrics, the existence of feedback classes, and polarity. Neither correctness, nor the performance, encouraging learners, or communicating knowledge gaps has been evaluated [10]. Subsequently,

the claims regarding LLM's capabilities have not been substantiated. Steiss et al. [56] included contextual feedback features, like providing directions for improvement, how accurate feedback has been, or whether essential feedback elements have been prioritized in their evaluation. On average, the human feedback performed better than the generated one[56]. Especially, non-accurate LLM-generated results have been criticized by the authors. They propose to combine auto-generated feedback with human ones. Feedback was not embedded into an educational setup. In another study, where formative feedback has been generated, the limitation of misleading information, and lacking information on task constraints was found [27]. The same challenges have been identified in more complex setups. Generated feedback also lacks when evaluating research papers [33].

While it can be helpful in some scenarios, an LLM is not capable of evaluating the methodological design or generating in-depth critique [33]. Tam [59, 60] identified that the quality of prompts used to generate feedback in an L2 language learning course highly influenced feedback quality. Stahl et al. [55] explored the correctness of LLM-generated scores for automatic essay scoring utilizing different prompt templates. To evaluate feedback quality, they focused on identifying mistakes, explanations, clearness, and suitability [55]. On these scales, the LLM had a high performance. Zhang et al. [67] identified that students perceive LLM-generated feedback positively in programming classes. Meyer et al. [38] analyzed LLM-generated feedback versus providing entirely no feedback and identified effects. Yet, the comparison to manually created feedback by teachers is still missing.

This paper explores the students, and teacher perspectives on LLM-generated feedback. To the best of the authors knowledge, there is a gap in research that investigate students' perceptions of LLM-generated versus teacher-generated feedback in academia. This paper aims to bridge that gap. Therefore, the following hypothesis are examined:

(1) There is a difference between students perception who get LLM-generated feedback to those who get manually created feedback.
(2) Students will primarily classify LLM-generated feedback correctly, identifying it as such, while recognizing manually created feedback as coming from the teacher.
(3) Students perceive feedback differently if they know it has been auto-generated.

3 Methodology

The study took place in a teacher education seminar for elementary school teaching at a German university, focusing on "the [German] language of schooling" [52]. The seminar's learning objective was to raise students' awareness of subject-specific and educational language requirements in the elementary school context, particularly in early education and literacy skill acquisition. A central task involved regularly documenting the learning process through specifically designed assignments that encouraged self-reflection on language-related beliefs and dispositions [14]. The task used for this study was the third of a total of five assignments, meaning that the students have already been familiar with the scope, objective, expected text product, and the type of feedback provided by the teacher. The assignment used for the study involved analyzing a short poem by Yoko Tawada ("Die zweite Person ich", i.e. "The Second Person I"), a Japanese-born but German-speaking poet, followed by a self-reflection on how the poem offers new perspectives on the German language and grammar. The seminar comprised four groups with identical content and tasks, involving a total of 32 participants, who gave their consent to analyze their responses. Two groups received authentic feedback, created by the teacher, while the other two received feedback generated by an LLM.

The feedback on these assignments was not intended to evaluate correctness or categorize responses as "right" or "wrong", particularly since there was no requirement to grade seminar performance. Instead, the feedback aimed to foster further intellectual stimulation, address attitudes that might impede learning, and clarify misunderstandings of concepts or positions. Since attitudes and dispositions, in particular, are deeply ingrained constructs of the human psyche and thus, represent one of the greater challenges in teacher education [17], the feedback process requires a highly appreciative approach: The learning steps already achieved must be acknowledged, and any critical positions must be explicitly recognized as understood, with elements that may still require reconsideration clearly framed as a personal enrichment for the future teaching profession and not as a personal deficiency or similar. The willingness to disclose personal opinions, thought processes, and sometimes very intimate attitudes toward socially sensitive issues must be explicitly acknowledged in the feedback process. Such feedback is aimed to be mimicked by an LLM.

First, an LLM capable of generating textual feedback has been deployed. Stahl et al. [55] identified that few-shot learning yielded the most helpful generated feedback. This approach leverages existing learner submissions along with teacher-generated feedback to effectively prime the LLM. Subsequently, learner submissions of varying quality can be integrated into the process. Building on this foundation, the paper employs a few-shot prompting strategy to incorporate historical student submissions and their associated feedback. This methodology is recommended as an effective means of contextualizing the LLM without necessitating fine-tuning [1]. Subsequently, previously known demonstrations can be incorporated so that the LLM-generated feedback can focus on similar aspects and uses the same tone as the teacher. To implement few-shot prompting, the LLM received historical essays with existing manually created textual feedback. The textual feedback included multiple aspects as described above.

Within the study setting, students are first asked whether the authors of this paper are allowed to make use of their submissions to train a feedback generator. Based on the resulting dataset, 10 submissions for which the students gave their permission to experiment with, are selected. More precisely, weak, regular, and high-quality submission are selected, which the teacher had rated in advance.

The prompting begins with a priming phase to clarify the LLM's role. Then, existing students' texts including feedback texts are combined to formulate few-shot prompts using system messages, followed by a prompt template. Table 1 shows the messages that have been designed for the study, refined through an iterative prompt engineering process, yielding encouraging results in the preliminary exploratory phase [50].

Table 1: Few-shot prompts to generate feedback.

ID	Role	Prompt
M1	system	You are a helpful teacher who provides feedback based on the texts submitted by students.
M2	system	Provide feedback on this text. Here are a few examples. Text1:"[student text1]", Feedback1:"[feedback text1]", Text2:"[student text2]", Feedback2:"[feedback text2]", ... Text10:"[student text10]", Feedback10:"[feedback text10]"
M3	user	Provide feedback on this text. Text: "[student text]" Maximum 250 words. Respond from the first-person perspective.

With that prompting, feedback can be generated for new student submissions. In a new semester, where the same seminar will be offered at the university, students are randomly allocated to two groups (1, and 2). Then, for a specific task, where students must write a texts, feedback is created. For group 1, the teacher manually creates typed feedback, which is comparable to previous semesters. Students of group 2 receive feedback that is generated using the few-shot prompted LLM. To address the hypotheses, the students finally get a questionnaire. This includes:

a) How helpful do you find the feedback of the last assignment? [free text]
b) I claim that 50% of the students received automatically generated feedback for the last assignment. Who provided the feedback? [single-choice options: from teacher, from LLM, uncertain]
c) How do you recognize whether the specific feedback comes from the instructor or from a tool? Provide examples from the feedback and justify your decision. [free text]
d) Let's assume the feedback was automatically generated. Now describe: How does this information change your perception of the feedback? [free text]

Based on the results of these questions, hypotheses 1)-3) can be answered. a) refers to 1), while student responses are clustered based on whether the feedback has been auto-generated, or not. Then, a coding schema is prepared, and student responses about the feedback on feedback are categorized employing that coding according to Mayring [36]. This is the basis to quantitatively compare identified categories among both student groups, employing Levene's test, and an independent samples t-test. 2) can be answered quantitatively by comparing the students' guesses, considering the source of the feedback (question b), and analyzing the students' correctness rates. This is comparable to a Turing test [62]. Furthermore, c) provides hints for correctly, or incorrectly classifying the feedback to be auto-generated, or not. Again, the coding according to Mayring [36] identifies indicators, for which the students claim to be able to distinguish auto-generated feedback from human ones. Levene's test, and an independent samples t-test for each indicator reveal which indicators lead to a correct identification of the feedback source. 3) can be answered qualitatively by d). Similar to the approach in 1), a coding schema is identified; however, in this case, to reveal positively, and negatively perceived aspects, which are then analyzed quantitatively. In order to avoid potential side effects, learners need to respond to an additional questionnaire to capture trait measurements about learning strategies [28]. These measurements are used to identify differences between both cohorts. Therefore, t-tests are conducted to explore potential side effects.

4 Results

We used Moodle as the learning management system, in which students submitted their texts. As described in the methodology section, from the historical text submissions, 10 students gave their consent to use their submissions, including the textual feedback that they received. Based on these texts, the few-shot prompt has been prepared (Table 1). The LLM Llama3.0:70b has been employed, which has been the state-of-the-art open model at the time of the experiment [37, 61]. In the new course, for the writing task, the Moodle activity "Task" has been used, and for the survey, the Moodle activity "Questionnaire". 60 students have been enrolled in the course. 50 of them completed the task, including the survey, from which 32 gave their consent to analyze their results. This set of individuals reflected a demographic distribution consistent with that of the entire seminar cohort. 16 students have been allocated to group 1, and 16 to group 2. From them, 16% were male and 83% female. Ages ranged from 20 to 50 years. 31% were in the 2nd semester, 5% were in the 3rd, and 64% were in the 4th semester. No one was in the first semester. A notable feature of the seminar and program was the inclusion of career changers (12%). Students focused on different teaching subjects. In the examined course, the following foci of students' subjects have been identified (multiple options have been possible): German (93%), general studies (73%), math (76%), special needs education (34%), sports (22%), and others 10%. In terms of learning strategies [28], no statistical significant differences have been identified between both groups.

4.1 Differences in students perception based on received feedback

Based on the initial qualitative analysis, four categories have been identified on which students' perceptions regarding the feedback they got can be classified:

(1) Whether the feedback provided has been perceived as helpful.

(2) Whether the students perceived the feedback as being motivational.
(3) Whether repetitions or reformulations of the students' submissions have been criticized.
(4) Whether content-related feedback has been emphasized by the students.

Notably, categories (1) and (2) pertain to students' perceptions of the feedback they received, addressing hypothesis (1), while the latter focuses on specific contextual limitations. Group statistics for these categories can be found in Table 2, while the learners were split based on from whom the feedback came from. It can be seen that there is no difference between the two groups when considering emphasized motivational aspects of the feedback. The auto-generated feedback has slightly been lower rated to be helpful, than the human-created versions. Furthermore, more content-related feedback was honored in the feedback that has been created by the teacher. Critique about repetitions can only be identified in the student group that got auto-generated feedback.

To identify whether statistical significant differences exist between the groups who received auto-generated, or manually created feedback, for each category, Levene's tests and t-tests were conducted. Table 3 summarizes the results. As identified in the group statistics, the motivational aspects do not differ among the considered groups. For the other categories, Levene's tests were significant with $p \leq .1$, especially for the "repetitive content" category, $P < .001$, indicating that equal variances cannot be assumed. Examining the two-sided significance of the Welch test (given that the assumption of homogeneity of variance cannot be assumed), $p = .02 < .10$ can be identified. Subsequently, there is a statistical significant difference under $p \leq .10$. The effect size is small to medium (Glass $\Delta = .479$). These results allow to draw conclusions for hypothesis 1. In terms of helpfulness, motivational aspects, or whether the feedback has been content-related, no statistical significant differences have been identified. However, there are differences in variances whether the feedback was perceived as helpful, or content-related. Repetitive aspects have been emphasized in the generated feedback only, as a prominent limitation of LLM-generated content. This finding is statistically significant, hence we can reject H_0, suggesting that there is a difference between the two groups considering repetitive feedback.

4.2 Correct classification of LLM-generated, or teacher-created feedback

Next, student's guess from whom the feedback comes from, is analyzed. Therefore, responses on question b) of the questionnaire, introduced in the methodology section are examined. Fig. 2 visualizes the number of students who got manually created feedback (blue bars), and those who got LLM-generated feedback (red bars) - in relation to their guesses. As the chart demonstrate, guesses are more or less equally distributed. Slightly more students who received feedback from the tutor also classified it to be manually created, than those who misclassified it. The same behavior can be seen for the LLM-generated feedback, for which slightly more students identified it as being LLM-generated. However, nearly the same amount is uncertain, who created the feedback. All in all, there is no statistical significant difference between both groups. Based on students who chose an option that the feedback came from the tutor, or that it has been LLM-generated (respectively, students who were uncertain, have been excluded), a final correctness rate could be derived, visualized in Fig. 3. 13 students predicted the correct source of the feedback ($\approx 59\%$), 9 students were not able to identify the correct source ($\approx 41\%$). There is no statistically significant difference whether the feedback has been manually created or LLM-generated. This answers hypothesis 2. In this experiment, students have not primarily been capable of classifying feedback as either LLM-generated or human-created. Furthermore, question c) identifies indicators for correct, or faulty conclusions. The following indicators have been identified: writing style (like gender-neutral language, use of emojis), circumstantiality, relationship to the student's text, described knowledge increase of the teacher, their shared experiences, generic paragraphs, overpraising, hallucinated content, and the time between the students submission and when they received the feedback. For each indicator, Levene's tests and t-tests are conducted to identify statistical significant differences between those who classified the feedback source correct, or incorrect. Levene's tests have been statistically significant for circumstantiality ($p = .003$), and overpraising ($p < .001$). Notably, the latter is also significant in the two-sided Welch test ($p = .04$), which has been employed instead of the standard t-test due to the violation of homogeneity of variance assumption as indicated by Levene's test. The effect size is small to medium (Glass $\Delta = .480$). Fig. 1 illustrates the indicators that the students described, based on whether they correctly identified the source of the feedback, or not. It can be seen, that only indicators about hallucinated content, overpraising, shared experiences by the teacher, and the timing have lead to correct allocations. The writing style has been a great indicator for $\approx 40\%$, but in $\approx 25\%$ of the cases, that indicator mislead the students' guesses. The others, especially circumstantiality, the relationship to the student's text, and the teacher's knowledge increase, also mislead the students in their guesses.

4.3 Students perceptions for auto-generated feedback

Lastly, d) is analyzed to identify students' perceptions for auto-generated feedback, as outlined in hypothesis (3). Based on the 32 textual responses, two main categories have been identified: content-related impacts and emotional/social effects. Each category has been subdivided along a tonal axis, giving rise to distinct positive and negative valences. The negatively content-related category includes notes about the assumption that feedback is less valuable, rather generic, and students have doubts about the auto-generated feedback including their submission. As positive/neutral impact, notes include responses like "Better than no feedback", or that students will not be affected. For the emotional/social effects category, different lacking elements within the feedback have been highlighted, which are negatively annotated, especially: the lack of interest/appreciation from the teacher, the lack of personal relationship, missing appreciation and satisfaction, disappointment as well as being demotivated when knowing that the teacher does not personally evaluate the student's submission. No positive emotional/social effects have been identified. Fig. 4 illustrates the relative number of responses. It can be seen, that the majority of

Table 2: Group statistics for feedback on feedback

Category	Feedback	N	Mean	Std. Deviation	Std. Error Mean
helpful	from teacher	16	.81	.40	.101
	auto-generated	16	.63	.50	.125
motivational	from teacher	16	.44	.51	.128
	auto-generated	16	.44	.51	.128
repetitive	from teacher	16	.00	.00	.000
	auto-generated	16	.31	.48	.120
content-related	from teacher	16	.38	.50	.125
	auto-generated	16	.19	.40	.101

Table 3: Levene's Test & Independent Samples Test (t-test/Welch test) on categories

		Levene's Test		t-test				
Category		F	Sig.	t	df	Significance Two-Sided p	Mean Difference	Std. Error Difference
helpful	Eva[1]	5.444	.027	1.168	30.000	.252	.188	.161
	Evna[2]			1.168	28.708	.253	.188	.161
motivational	Eva	.000	1.000	.000	30.000	1.000	.000	.181
	Evna			.000	30.000	1.000	.000	.181
repetitive	Eva	91.667	<.001	-2.611	30.000	.014	-.312	.120
	Evna			-2.611	15.000	.020	-.312	.120
content-related	Eva	5.444	.027	1.168	30.000	.252	.188	.161
	Evna			1.168	28.708	.253	.188	.161

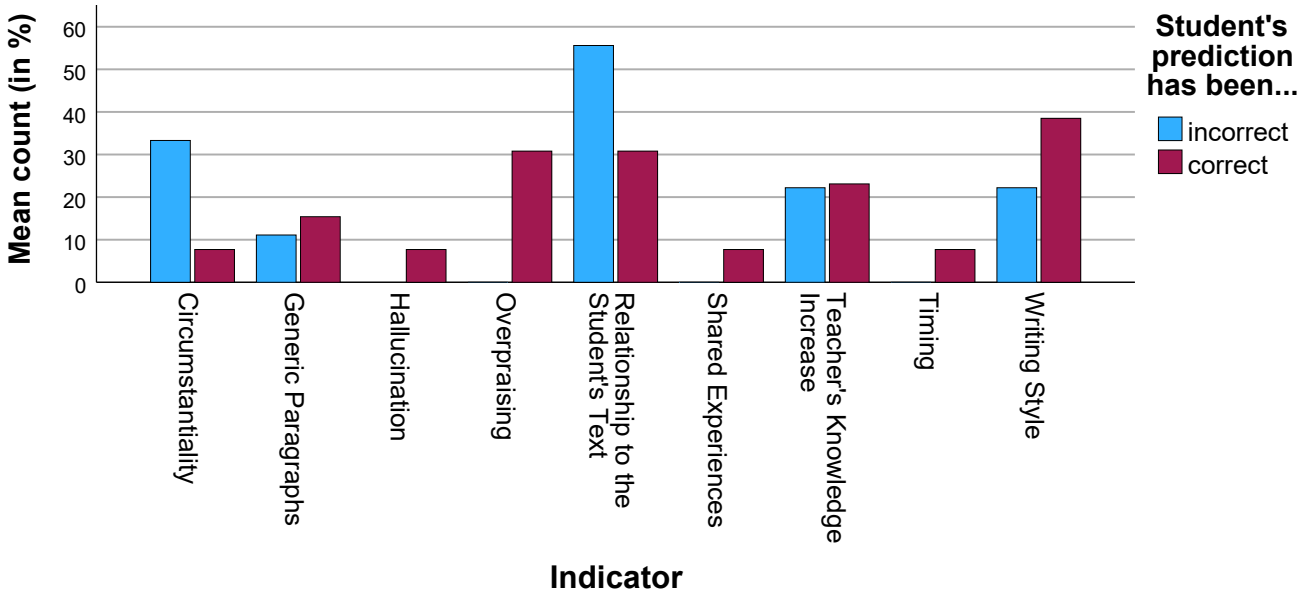

Figure 1: Students' indicators for predictions, partitioned by correctness in prediction.

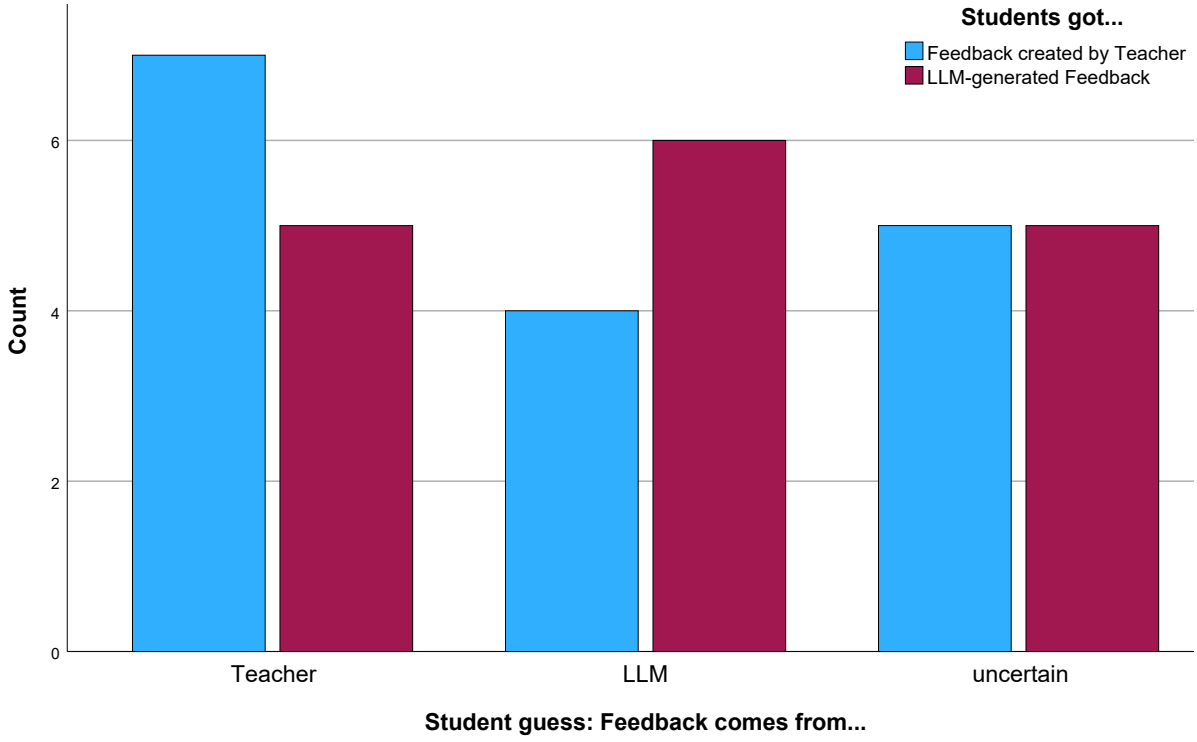

Figure 2: Students guess, from whom the feedback comes from for group 1 (teacher wrote feedback, blue), and group 2 (LLM generated feedback, red).

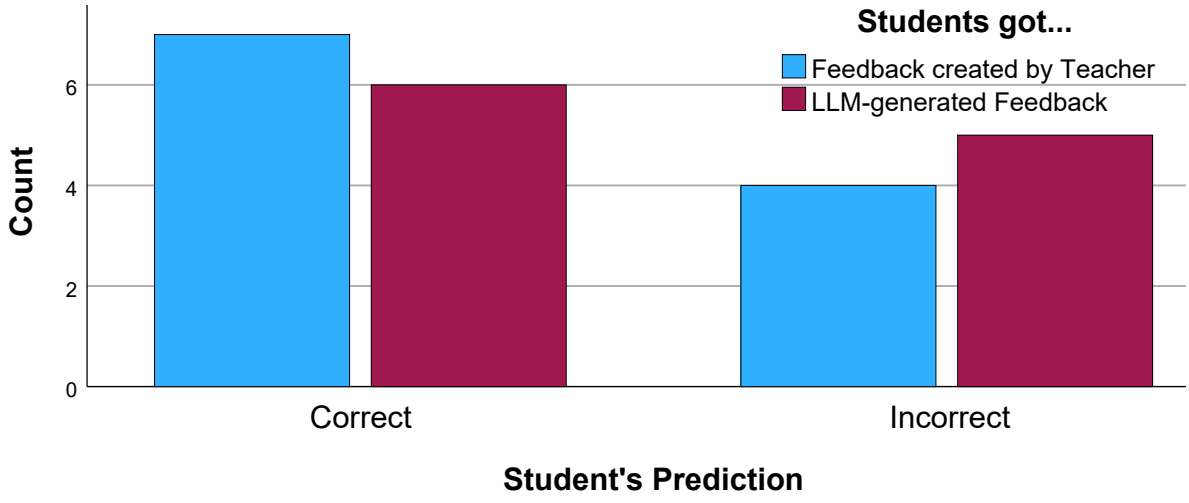

Figure 3: Number of students who classified correct, or incorrect, and which type of feedback they received (LLM-generated or manually created by teacher).

students perceives auto-generated feedback negatively. The number of negatively perceived content-related concerns is similar to negatively perceived emotional/social effects. From the content-related impact, 62.5% of the students expect, that the feedback is less valuable, which is the largest class found in that category, followed by doubts (9.4%). Only one student reported both classes, hence the final total number equals 68.8%. Within the emotional/social effects, 43.8% emphasize the lack of interest/appreciation from the teacher, which is also the largest class in that category, followed by the lack of personal relationship (6.3%). This result contradicts the

findings of 4.1. In Table 2, it can be seen that in 63% of the cases, auto-generated feedback has been rated as helpful. The mean count for positively perceived content-related aspects have been slightly similar if all students have been asked (15.6%, Fig. 4), in relation to the focus on students who got auto-generated feedback only (19%, Section 4.3, Table 2). The result shows that content-related aspects of real-world LLM-generated feedback align with students' perceptions. In Section 4.3, Motivational aspects have been the same, independently from the feedback's source. However, when asking students how they perceive auto-generated feedback as done, it has been rated as being highly demotivating, without any positive emotional/social effects. Subsequently, motivational aspects, which are part of the emotional/social effects, are highly undermined when knowing that the feedback has been auto-generated (0%, Fig. 4). This contradicts the result in the real-world setting for motivational aspects (44%, Table 2). Subsequently, hypothesis 3 can be answered.

To summarize, students perceive the same number of positive, content-related aspects in auto-generated feedback, whether they are aware it is auto-generated or not. Considering motivational aspects, students perceive auto-generate feedback differently if they know it has been auto-generated.

5 Discussion

The study has shown that students have not been capable of identifying, whether feedback was generated by an LLM, or a teacher. Even though, generated feedback was highly negatively perceived when students know that it has been generated. The relational aspect highlighted in the study is particularly interesting. The examination has demonstrated that a gap between students and the teacher can arise, especially if the concept of auto-generated feedback has not been sufficiently clarified. However, such a disruption of the pedagogical relationship is a significant obstacle, especially in feedback processes, as feedback requires "a climate of high trust and reduced anxiety" [19]. This limitation, however, only arises under the condition that the feedback has been auto-generated, as the same feedback may otherwise perceived as helpful when provided by a teacher (Table 2). This reveals an apparent (emotional) qualitative difference in whether the submitted work has been reviewed and praised by a real human or an LLM, with the human praise being clearly valued more. For educational contexts, this means that the use of LLMs in feedback processes must be carefully explained. Particularly in tasks that do not have a specific sample solution but instead require the formation of opinions, personal positioning, or the disclosure of personal attitudes, it is essential to maintain the human element in the feedback process. This should be done transparently to avoid undermining supportive relational structures in educational settings.

Additionally, more research is needed on actions that are taken (or not taken) in response to received feedback: "That students are taught to receive, interpret and use the feedback provided is probably much more important than focusing on how much feedback is provided by the teacher, as feedback given but not heard is of little use" [19]. Subsequently, when redesigning a feedback generator, it is important to communicate "the essence" only, based on a small set of feedback criteria. A feedback generator could identify a criteria set, from which the teacher can choose the essential ones. Then, the selection can be based on a pedagogical baseline, which the teacher is responsible for. However, in general, as the feedback is text-based, it can still be optimized by the teacher. The experiment revealed that especially hallucination, and overpraising have been typical for LLM-generated feedback. Also, there is the need for more content-related feedback (Table 2), which the teacher may add. Subsequently, the feedback generator can be employed to generate the foundational feedback text, which the teacher may then refine, enhance with additional aspects, or amend as needed. Then, especially writing generic generic parts can be done automatically. Hence, teachers may focus on further aspects to create high information feedback, with regard to the pedagogical essential aspects. In addition, teachers must be aware of how to use auto-generated feedback without blindly trusting the output. With this obstacle, teachers should have an understanding how feedback generation with LLMs principally works. Misconceptions about LLMs' capabilities for feedback generation should especially be addressed in teacher education to enable educators to successfully use such tools. As a consequence, much more pedagogical work is required on feedback literacy, particularly regarding possible feedback sources (whether auto-generated or human-created), with a greater focus on individuals' own feedback capacities, especially for those in pedagogical professional contexts. Therefore, it is essential not only to increase the number of feedback situations and clarify them but also to train future teachers to use tools like LLMs, e.g., in peer feedback. Referring to the concept of feedback literacy, the study of the paper reveals a clear need for further research, particularly on how LLMs affect the development of feedback literacy among educators. For instance, competence models may need to be expanded or adapted in terms of the use of LLMs for feedback purposes [2].

However, the perception about generated feedback does not need to be limited to the students perspective. Also, teachers may benefit from it. When the generator has been trained with historical feedback, it mainly mimics the teacher's style, and tone. It also mimics the structure of the historical feedback. Teachers can reflect their feedback when being confronted with mimicked, and reproduced versions. Especially focusing on positive aspects and already identifiable resources in student responses, which is often neglected due to time constraints and tends to lead to more negative feedback, can be brought back to the attention of teachers through the use of generators. Thus, the feedback generator can also be a great tool for self-evaluating the own feedback practice.

Future feedback, employing prompts with LLMs, should be more task-specific, targeting learning goals, and fostering independent learning processes. A next step could involve measuring potential learning outcomes: Do students who receive automated feedback perform as well as those who receive authentic, human-generated feedback? By assessing performance and outcomes through various metrics, such as grades, self-reflection, and the quality of work, you could explore whether the effectiveness of automatic feedback aligns with or diverges from that of traditional feedback. An additional optimization could involve explicitly asking students how they plan to use the feedback in their learning process and what their first steps will be. In the context of teacher education, focusing on how feedback is practically utilized can lead to valuable insights

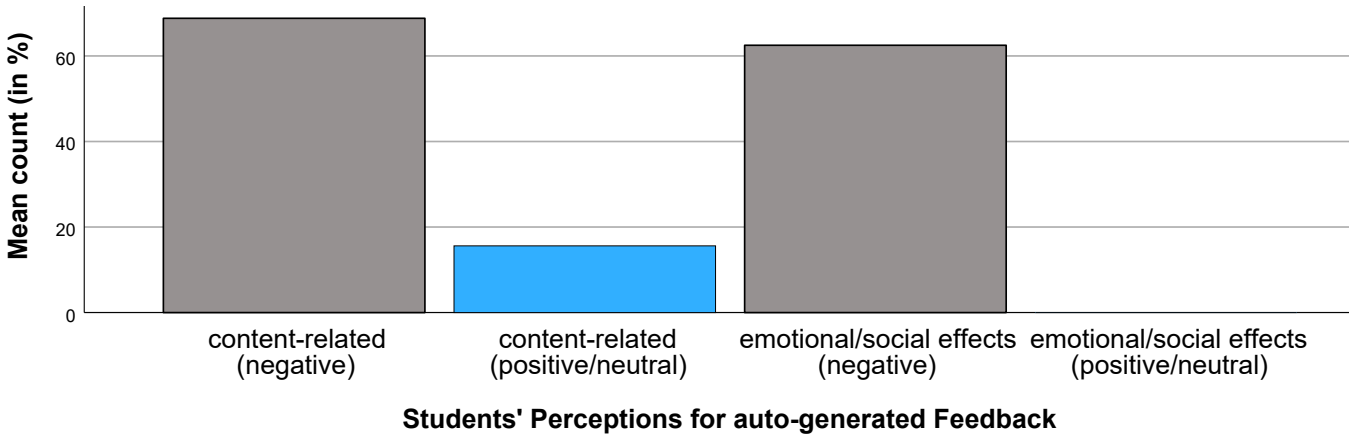

Figure 4: Relative Number of Responses by categories, and polarity.

for future teaching practice. When feedback encourages reflection and further development, it helps prospective teachers to incorporate these practices into their future classrooms. By integrating suggestions from feedback into their own learning process, student teachers can better understand how to foster self-regulated learning and continuous improvement in their future students. This reflective approach thus can bridge feedback and real-world applications in education.

An essential limitation is the use of LLMs as black box technologies in education. Although we employed Llama, an foundational open-source LLM, there is no functionality to explain indicators for generated feedback. However, certain indicators representing feedback criteria, are missing. In this paper, samples of historical feedback have been used for prompting. Aiming to generate explainable feedback, the capabilities of LLMs could be taken to extract results of a pre-defined set of feedback criteria first. Utilizing a criteria grid is a typical procedure when evaluating student submissions to provide feedback, for example based on language proficiency levels according to the CEFR in language learning [43, 58]. The LLM's responses for each criterion must be evaluated to identify those that align with teacher expectations [50]. In combination, a set of explainable feedback indicators could be derived, which can then be used either to directly formulate feedback, or if the grid of feedback criteria already covers explanations, they can directly be copied to formulate feedback, based on several fitting text snippets. With that approach, the capabilities of LLMs would be used, but feedback can be generated with explanations. In a more sophisticated view, feedback indicators does not need to be limited to direct explanations. Based on LLM-generated ratings for feedback indicators, and historical teacher ratings, supervised machine learning models can be trained, that allow to predict ratings of the feedback criteria grid. With such an approach, combinations of feedback indicators could automatically be learned, so that optimal feedback text snippets can then be selected [49]. This can be the scope of future work.

In the experiment, prompt engineering has been limited to few-shot learning, employing a set of historical learner submissions, including the feedback, which they got. This approach has the advantage, that if students withdraw the permission to use their submissions, the prompt can easily be modified, as the entire set of submissions is part of the prompt. However, LLMs are often sensitive to changes in prompting. If the prompt changes, the response may be completely different. In general, the behavior of LLMs can be controlled, like the temperature which represents the degree of variances in responses. For the paper, the prompt with the most promising result has been used, but there might be other prompts leading to significantly better results. Furthermore, employing few-shot prompting leads to another limitation. As the number of tokens per prompt is limited, the approach of priming the LLM can be limited by its length. Subsequently, depending on the lengths of learners' submissions, including feedback texts, only a reduced number can be included. In the case of allowing a wide range of topics to be included in the assessment, only a subset may be covered. Nevertheless, the prompting results have already been promising, despite the limitation of the length of historical submissions. The state-of-the-art LLMs, with the designed prompts, can only handle roughly one page. Evaluating thesis, or longer texts can currently not be processed, only portions of them. Nevertheless, when historical students' submissions including textual feedback are available, and if we have the permission to use them, the approach of LLM-generated feedback can be employed. In this experiment, feedback on self-reflection tasks in a teacher education course for further intellectual stimulation is used. It is expected, that for similar tasks, the generator will produce similar results. However, further research should examine whether the promising results also arise beyond the domain of teacher education for elementary school education, and for other task types.

The integration of automated tools can facilitate self-assessment exercises by providing students with prompts that stimulate reflection on their work, informed by generated feedback. Moreover, such tools may offer particular value in situations where students are hesitant to share personal reflections related to their future professional roles as teachers, and still seek constructive feedback. In this case, a well-programmed generator may be leveraged to facilitate learning. However, it is crucial to explicitly acknowledge and convey the lack of human teacher input within the feedback process.

Additionally, feedback generators could be used for evaluating lesson plans, analyzing self-created teaching materials, and assessing the difficulty levels of language-related hurdles in educational content. This broadens their applicability in teacher education.

Despite the flexibility of LLMs, there is the difficulty of integrating them into software components [64]. For the study, the feedback generator has been implemented as a separate tool, which the teacher accessed. As only one prompt is automatically created, handling a single request, and process the response, which includes the entire generated feedback, is operationalizable. However, if more complex approaches will be implemented that includes multiple prompts, that need to be combined, the output formatting need always to be correct to process gathered information in a pipeline-like architecture. For a ready-to-use software, a robust architecture with error handling, and automatisms, is required. At the time of the study, the feedback generator has been a software prototype. As the scope of the paper has been to address the three perspectives as introduced in the related work section, implementing a robust software architecture has been omitted. This can be the scope of future work.

6 Conclusion

In this paper, students' perceptions regarding LLM-generated feedback has been investigated. The results revealed, that:

(1) Repetitive content in LLM-generated feedback has been criticized, revealing a statistically significant difference compared to students who received teacher-based feedback. However, no statistically significant differences were identified in terms of motivational aspects or perceived helpfulness.
(2) Students have not been capable of correctly identifying the feedback source. A small set of indicators has been found, which can identify the feedback source, namely hallucinated content, overpraising, shared experiences by the teacher, and feedback submission time.
(3) Students highly criticize and reject auto-generated feedback. Negative aspects indicate that students perceive emotional and social effects, in particular, as being undermined when feedback is auto-generated. This highly contradicts the finding in (1), where no differences for motivational aspects could be found.

From the pedagogical perspective, there is an essential take-home-message: The technology can beneficially be employed, especially when students seek first feedback. The teacher can use LLMs to create a first feedback draft, which they can optimize. As the generator employing few-shot promoting makes use of the teachers writing style, the generated feedback is a great basis to create high-information feedback. Nevertheless, it is essential to keep the teacher in the loop. They need to be aware to have a look at the feedback to remove hallucinated content, or faulty conclusions, and to highlight aspects which they have identified to be important for the learning process. LLMs can be supportive, but cannot replace teachers. From the teacher's perspective, self-reflective insights can also be drawn from this work. As shown, students have gaps in their knowledge about the capabilities and limitations of LLMs. More clarity and transparency are needed for the future use of such generators in higher education, including the teacher's role in staying in the loop, to avoid negative impacts on relationships. Student teachers should also independently experiment with feedback tools to explore their capabilities. This could help mitigate the negative perceptions associated with automated feedback. Further research will show how adaptable the promising feedback generator will be in other domains.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DHBKI045.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[2] Simon Buckingham Shum, Lisa-Angelique Lim, David Boud, Margaret Bearman, and Phillip Dawson. 2023. A comparative analysis of the skilled use of automated feedback tools through the lens of teacher feedback literacy. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 40.
[3] David Carless. 2006. Differing perceptions in the feedback process. *Studies in higher education* 31, 2 (2006), 219–233.
[4] David Carless and David Boud. 2018. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education* 43, 8 (2018), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354
[5] Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. 2023. MultiPL-E: a scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering* (2023).
[6] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.
[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
[8] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044* (2019).
[9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).
[10] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 323–325.
[11] Oxford English Dictionary. July 2023. feedback, n., additional sense. https://doi.org/10.1093/OED/6371229855
[12] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems* 36 (2024).
[13] Catherine Garrison and Michael Ehringhaus. 2007. Formative and summative assessments in the classroom.
[14] Axinja Hachfeld. 2011. Lehrerkompetenzen im Kontext sprachlicher und kultureller Heterogenität im Klassenzimmer : Welche Rolle spielen diagnostische Fähigkeiten und Überzeugungen?
[15] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints* (2023).
[16] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* (2023).
[17] Arran Hamilton and John Hattie. 2022. *The lean education manifesto: a synthesis of 900+ systematic reviews for visible learning in developing countries.* Routledge.
[18] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for

adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509* (2022).
[19] John Hattie and Shirley Clarke. 2018. *Visible learning: feedback.* Routledge.
[20] John Hattie, Mark Gan, and Cameron Brooks. 2016. Instruction based on feedback. *Handbook of Research on Learning and Instruction (2nd ed.). Routledge. https://doi.org/10.4324/9781315736419* (2016).
[21] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.
[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
[23] Ana Isabel Hibert. 2019. Systematic literature review of automated writing evaluation as a formative learning tool. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14.* Springer, 199–212.
[24] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2023. Large language models for software engineering: A systematic literature review. *arXiv preprint arXiv:2308.10620* (2023).
[25] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.
[26] Jonathan D Kibble. 2017. Best practices in summative assessment. *Advances in physiology education* 41, 1 (2017), 110–119.
[27] Natalie Kiesler, Dominic Lohr, and Hieke Keuning. 2023. Exploring the potential of large language models to generate formative programming feedback. In *2023 IEEE Frontiers in Education Conference (FIE).* IEEE, 1–5.
[28] Katrin B Klingsieck. 2019. Kurz und knapp–die Kurzskala des Fragebogens „Lernstrategien im Studium "(LIST). *Zeitschrift für pädagogische Psychologie* (2019).
[29] Alison C Koenka, Lisa Linnenbrink-Garcia, Hannah Moshontz, Kayla M Atkinson, Carmen E Sanchez, and Harris Cooper. 2021. A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology* 41, 7 (2021), 922–947.
[30] Shaofeng Li. 2010. The effectiveness of corrective feedback in SLA: A meta-analysis. *Language learning* 60, 2 (2010), 309–365.
[31] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance.* 374–382.
[32] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).
[33] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* (2024), AIoa2400196.
[34] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804* (2021).
[35] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics.* Springer, 387–402.
[36] Philipp Mayring. 2004. 5.12 Qualitative Content Analysis. *A companion to qualitative research* (2004), 266–269.
[37] AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date, 2024. *URL https://ai. meta. com/blog/meta-llama-3/. Accessed on April* 26 (2024).
[38] Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 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 (2024), 100199.
[39] Bushra Mohammad, Turjana Supti, Mahmood Alzubaidi, Hurmat Shah, Tanvir Alam, Zubair Shah, and Mowafa Househ. 2023. The pros and cons of using ChatGPT in medical education: a scoping review. *Healthcare Transformation with Informatics and Artificial Intelligence* (2023), 644–647.
[40] Emma Mulliner and Matthew Tucker. 2017. Feedback on feedback practice: perceptions of students and academics. *Assessment & Evaluation in Higher Education* 42, 2 (2017), 266–288.
[41] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality* 15, 2 (2023), 1–21.
[42] David J Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education* 31, 2 (2006), 199–218.
[43] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge University Press.
[44] Maciej Pankiewicz and Ryan S Baker. 2023. Large Language Models (GPT) for automating feedback on programming assignments. *arXiv preprint arXiv:2307.00150* (2023).
[45] Lorena Parra G and Ximena Calero S. 2019. Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction* 12, 2 (2019), 209–226.
[46] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193* (2021).
[47] Sylvio Rüdian, Moritz Dittmeyer, and Niels Pinkwart. 2022. Challenges of using auto-correction tools for language learning. In *LAK22: 12th International Learning Analytics and Knowledge Conference.* 426–431.
[48] Sylvio Rüdian and Niels Pinkwart. 2023. Auto-generated language learning online courses using generative AI models like ChatGPT. (2023).
[49] Sylvio Rüdian, Clara Schumacher, Jakub Kužílek, and Niels Pinkwart. 2023. Preselecting Text Snippets to provide formative Feedback in Online Learning. In *Proceedings of the 16th International Conference on Educational Data Mining.* 430–433.
[50] Sylvio Rüdian. 2024. *Exploratory and Confirmatory Prompt Engineering.* Berlin, DE, 1–6. https://doi.org/10.5281/zenodo.12549309
[51] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM* 64, 9 (2021), 99–106.
[52] Mary J Schleppegrell. 2001. Linguistic features of the language of schooling. *Linguistics and education* 12, 4 (2001), 431–459.
[53] Ken Sheppard. 1992. Two feedback types: Do they make a difference? *RELC journal* 23, 1 (1992), 103–110.
[54] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189.
[55] Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. *arXiv preprint arXiv:2404.15845* (2024).
[56] Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction* 91 (2024), 101894.
[57] Marie Stevenson and Aek Phakiti. 2014. The effects of computer-generated feedback on the quality of writing. *Assessing Writing* 19 (2014), 51–65.
[58] Christine Tagliante. 2016. *L'évaluation et le CECR. Techniques et pratiques de classe.* Clé International.
[59] Angela Choi Fung Tam. 2024. Interacting with ChatGPT for internal feedback and factors affecting feedback quality. *Assessment & Evaluation in Higher Education* (2024), 1–17.
[60] Qiuyu Tao, Jiang Zhong, and Rongzhen Li. 2022. AESPrompt: Self-supervised Constraints for Automated Essay Scoring with Prompt Tuning.. In *SEKE.* 335–340.
[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[62] AM Turing. 1950. Computing machinery and intelligence, The Turing Test: Verbal Behavior as the Hallmark of Intelligence. (1950).
[63] Fabienne M Van der Kleij, Remco CW Feskens, and Theo JHM Eggen. 2015. Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research* 85, 4 (2015), 475–511.
[64] Irene Weber. 2024. Large Language Models as Software Components: A Taxonomy for LLM-Integrated Applications. *arXiv preprint arXiv:2406.10300* (2024).
[65] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).
[66] Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology* 10 (2020), 487662.
[67] Zhengdong Zhang, Zihan Dong, Yang Shi, Thomas Price, Noboru Matsuda, and Dongkuan Xu. 2024. Students' perceptions and preferences of generative artificial intelligence feedback for programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23250–23258.
[68] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
[69] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364* (2023).