Didaktik Deutsch

Maurice Fürstenberg

Measuring the quality of AI-generated feedback

From theoretical modelling to empirical evidence

Abstract

This study takes a theoretical and empirical approach to exploring the quality of AI-generated feedback on texts. First, it reviews the current state of research and demonstrates challenges involved in operationalizing feedback quality. Furthermore, it develops a theoretical three-level model that establishes the necessary terminology. The study then compares how 75 experienced teachers and three AI systems (Mistral Large, Llama 4 Maverick, GPT-4.1) provide feedback on three pupils' texts. The feedback takes the form of criterion-based scores, overall grades, and short feedback texts. Eight trained raters assessed the quality of the feedback texts based on set criteria.

The teachers and AI systems (across 10 iterations) showed high interrater reliability (*ICC* = 0.7–0.9). The AI models consistently gave higher grades, but in the same ranking order as the teachers. The criterion-based scores differed significantly. Teachers weighted the individual criteria independently when predicting the overall grade, whereas AI systems did not. The strongest predictors of the quality of the feedback texts were concreteness and explanation. However, the source (AI vs. teacher) had no significant influence on the perceived usefulness.

Keywords: Writing • AI • Automated Writing Evaluation • Large Language Models • Feedback

Didaktik Deutsch
Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur
Special Issue (2026) – Writing with AI. S. 65–96
DOI: 10.21248/dideu.966
Copyright Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0
veröffentlicht: https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de

1 Introduction

The increasing use of generative AI[1] systems as "tutor" (Fürstenberg & Müller, 2024) or "writing tutor" (Steinhoff, 2025) in, for example, automated text feedback presents significant opportunities and challenges for research on teaching and writing. Given the transformative potential of AI (Kasneci et al., 2023), many questions arise about the extent to which established writing practices can or must be changed or even replaced. In particular, the use of AI-generated systems as feedback machines has attracted significant interest since 2023 (Kaliisa et al., 2025). This interest is hardly surprising, given that feedback is of central importance for learning how to write (Graham et al., 2017; Hattie & Timperley, 2007; Skar et al., 2022). Yet, learners often receive it in insufficient quantities and heterogeneous quality (Müller et al., 2023). Creating formative feedback (Philipp, 2023) is a learning process for teachers themselves (Sturm, 2016) that requires considerable effort and is a very time-consuming task (Henderson et al., 2019; Mußmann et al., 2016, 2020), whereas AI systems such as ChatGPT, Llama, or Mistral can produce feedback in no time at all.[2] Furthermore, the German education system is currently facing a significant shortage of teachers, and this problem will foreseeably continue in the future (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik, 2022). One problematic consequence of this situation is that AI-based applications are widely used (as feedback machines) in the educational landscape (Limpert, 2025). This means that a considerable amount of public money and, in some cases, even teachers' private budgets are invested in these systems without much consideration of their quality, usefulness, or consequences for pupils and teachers.[3] In principle, AI systems do not understand what they are doing (Fürstenberg & Müller, 2024). They are "language use machines" that simulate understanding (Müller & Fürstenberg, 2023). Among other things, the quality of AI-generated feedback is therefore questionable.

2 Related work

This section establishes the theoretical foundation of the study. First, the concept of feedback is defined. Then, to systematize studies on feedback quality and clarify terms, a three-level model is introduced. Subsequently, a brief overview of current research on the quality of AI-generated feedback is provided along with an analysis of research gaps.

2.1 Conceptual and methodological challenges in evaluating AI-generated feedback

Following Hattie and Timperley (2007), feedback is understood as information that is communicated to learners with the aim of closing a gap between an actual and a desired performance. This information can be given at various stages of the development of (writing) competence (Christophel et al., 2017) and by different agents (Panadero & Lipnevich, 2022). The state of research on AI-generated feedback is already vast and difficult to comprehend (Fong, 2025; Kaliisa et al., 2025). This is partly

[1] Currently, the term "AI" seems to be undergoing a semantic narrowing (Fürstenberg & Müller, 2024) to refer specifically to the concept of "generative LLM", a convention that is also followed in this article. Therefore, when AI is mentioned below, it usually refers to generative LLMs.

[2] For a critical and historical perspective on the automation of teachers' work, see Rensfeldt and Rahm (2023).

[3] The perspective of (prospective) teachers has largely been overlooked in research. It remains unclear what impact the use of AI systems (as feedback machines) will have on the self-concept/-esteem/-efficacy/-awareness, motivation, and especially the skills of (prospective) teachers. Lee et al. (2025) indicate that high time pressure and low skills lead to a lack of critical thinking when dealing with the products of AI systems. This poses significant risks for the education sector that must be critically monitored by researchers.

because feedback is naturally studied in all areas in which learning takes place.[4] However, the present study focuses on the quality of feedback on (mostly language) learners' writing products provided by AI systems. Therefore, this section provides an overview of how previous studies have approached the question of the quality of AI-generated feedback on learners' writing products. A further reason for focusing on methodology instead of presenting results becomes clear in the following: The higher-level question concerning feedback quality (i.e., how good is [AI-generated] feedback?) involves so many decisions at the methodology level with regard to the operationalization of central concepts that a presentation of the results or even a summary would be out of place here. The studies differ greatly in their answers to questions such as the following: In what form is feedback delivered, and how was it generated? In what modality is feedback delivered? Who creates feedback and who delivers it? How skilled is a feedback giver on the task/feedback? When is feedback given? How often is feedback given? How is feedback formulated? What is the focus of feedback? What is the scope of feedback? What is the goal of feedback? Furthermore, questions about students receiving feedback can be summarized under a higher-order question: Who receives feedback? Subsequent questions could concern, among others, their skills, motivation (Jansen, Höft, et al., 2025), strategies, personality (Meyer et al., 2025), age, and cultural context (Fokides & Peristeraki, 2025). All these and other questions can be answered with different responses. Depending on how each question is answered, a completely different picture of what constitutes good feedback emerges. Most of these questions and answers apply equally well to studies that examine feedback quality without the use of AI. The integration of AI in these studies is further complicating the comparability of the study's results by the changing (some say "evolving") models and their probabilistic nature; as well as by specific methodological decisions such as the design of the ((foundation) system) prompt (Neumann et al., 2025), the temperature settings, or simply the size of the models in terms of the number of parameters learned during training or their context window size (Jung et al., 2025). This challenge will grow even further with each new study that takes a different approach, and it shows that a section such as the present one would be premature, if it was presenting results without reflecting the specific methodology.

2.2 A three-Level model of (meta-)feedback in empirical research

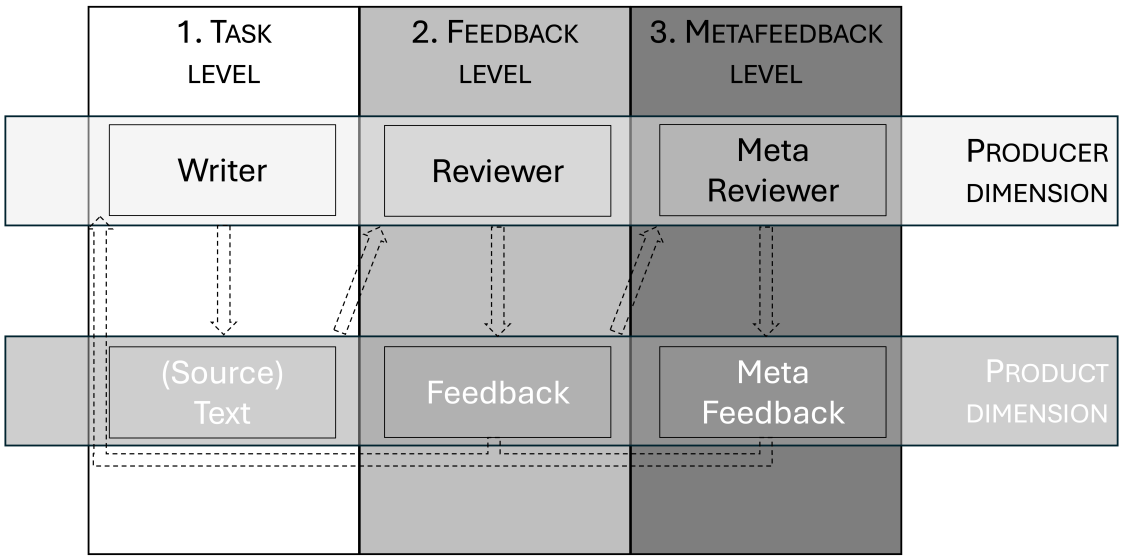
For terminological clarity, a theoretical model will systematize the methodology used in related studies. If examining research on feedback, it is impossible not to notice that it can quickly become confusing: The concept of feedback alone implies that at least two levels are integrated: Someone or something must produce something to which any kind of feedback can be given by somebody or something. A lot of studies now establish a third level, because they look at this two-level process (Fong, 2025) from a metaperspective (e.g., Banihashem et al., 2024; Dai et al., 2024; Jansen et al., 2024). Therefore, these studies implement any kind of one or more producers to generate any kind of metafeedback on the feedback itself. The resulting metareviewer/-s is/are essentially an operationalization compromise[5] to get closer to the construct of feedback quality. In order to systematize the various producers and products involved in the feedback studies, Figure 1 provides an overview:

[4] Incidentally, it would be a promising project to consider these different perspectives on these cross-cutting themes of feedback and AI.

[5] Although it might be assumed that this is a phenomenon limited to scientific studies, metafeedback is actually implemented in (German) school practice: for example, in the training of preservice teachers in which instructors give teachers metafeedback; or for quality assurance purposes in schools in which teachers in special functions check the corrections in tests from other teachers in a second correction and also communicate this back to their colleague.

Figure 1

Systematic View of Different Levels and Dimensions in the (Meta) Feedback Process



There are three levels and two dimensions: On every level, there are different types of producers (= producer dimension) and mostly texts (= product dimension) that are involved in the majority of the feedback studies, each of which can be placed in one dimension. For producers, there are the writers, the reviewers (i.e., feedback providers), and the metareviewers (i.e., feedback assessors). In the dimension of products, there is the (source) text from the students to solve a given task, the feedback (which is mostly a text too, but can also be points on criteria, a grade, or corrections), and often also metafeedback that can also be text but must not be (see below). On the producer dimension as well as on the product dimension, there can be more than one producer or product (e.g., the feedback consists of a text and points on criteria). The arrows show a typical path of this recurrent process that normally begins with the writer. Some revisions of the source text are also available in studies (as in Li et al., 2024; Meyer et al., 2024; Zou et al., 2025). These are taken into account here by the arrow below.

The three levels on which products and producers operate can be chronologically systematized as follows with regard to the central phenomenon (feedback):

1. Task level: level of what feedback is given for
2. Feedback level: level of the feedback itself
3. Metafeedback level: level of the assessment of the feedback[6]

Writer, reviewer, and metareviewer can, but must not, be the same person/agent. Especially from writer to reviewer and from reviewer to metareviewer, most of the studies change the producer. The default situation in schools is that a pupil is the writer and normally also an (implicit) metareviewer (if they read the feedback) as seen in Rüdian et al. (2025). The reviewer is normally the teacher but can

[6] Sometimes the metafeedback level does not meet the producer dimension, because the operationalization of feedback quality did not include another producer to feedback the feedback, but quantitatively compared two different reviewers' points (e.g., human vs. human, AI vs. AI, or human vs. AI). Nevertheless, it seems permissible to keep the model in Figure 1, because this kind of operationalization just skips the person level to give metafeedback directly (e.g., in the form of statistical interrater reliability values). This is also the reason for the term *product* (as a statistical value is a product) instead of text.

also be part of the peer group (Banihashem et al., 2024; Shin et al., 2025; Weidlich et al., 2025)—for example, as part of writing conferences (Fix, 2004; Lehnen, 2023; Reichardt et al., 2022). In contrast, a prototypical study looks like this: University students or secondary school pupils (writers) write a text (source text) and receive feedback from a person and an AI model (reviewer). This feedback is then evaluated in a questionnaire (metafeedback) by trained raters (metareviewer) based on various criteria for the quality of feedback. From a school's perspective, this comparison highlights an initial issue (ecological validity) with some of the studies, as it is difficult to compare, for example, university students and school pupils. Such sampling biases in educational research are well documented in psychology (Arnett, 2008) and has since been shown to affect applied linguistics as well: Andringa and Godfroid (2020) found that 88% of all adult samples in the studies they examined consisted of university students. I have observed a similar pattern in the studies reviewed here and suspect that this bias is further reinforced by the term *student*, which in English can refer both to university students and to school-age learners — a distinction that German marks unambiguously with *Studierende* and *Schüler*. To avoid this ambiguity, I consistently use the term *pupils* when referring to secondary school learners in this study. The following provides a brief overview of 32 studies.

2.3 Related studies: A brief overview

To identify relevant studies, a keyword search ("feedback," "AI," "writing," "students," "teachers," "automated essay scoring," "automated writing evaluation") on *Google Scholar*, *arxiv*, and the databases of the *Association for Computational Linguistics* (*ACL*) and the *Association for Computing Machinery* (*ACM*) was conducted as well as a snowballing approach. Studies that did not address feedback quality, texts, or language learning in any way, or that had not undergone a double-blind peer review process were excluded. Based on the current status 32 studies remained.[7] The methodology of these studies is summarized below.

Across the 32 reviewed studies, which are all published in English-language journals, feedback was provided on approximately 10,441 learner texts. The average age of participants was 19 years. In total, 10 studies involved pupils from schools, whereas the remaining studies were conducted at the university level. This is understandable given the central processes involved in data collection (approvals from authorities and participants, ethical issues); something that often involves significantly more barriers in schools. The L1 of the participants is mixed: Six studies examined German students' texts, three of which focused on texts written by secondary school pupils (Jansen et al., 2024; Schaller et al., 2024; Seßler et al., 2025). All studies focused on texts as the central learning objects; most of the studies (13) addressed English as a Foreign Language (EFL) writing.

In 13 cases, feedback quality was purely operationalized in a quantitative manner, typically within the framework of automated essay scoring (AES) approaches comparing human and large language model (LLM) reviewers. In 11 additional studies, feedback quality was assessed through combined operationalizations such as learner self-reports, analyses of text revisions, metarater agreement on criteria of good feedback, content analyses, or automated linguistic evaluations. With 3 exceptions (Rüdian et

[7] Almegren et al. (2024); Alsofyani and Barzanji (2025); Aydın et al. (2025); Banihashem et al. (2024); Bannò et al. (2024); Bui and Barrot (2025); Chiang and Lee (2023); Dai et al. (2024); Escalante et al. (2023); Fokides and Peristeraki (2025); Guo and Wang (2024); Jansen et al. (2024); Kim (2025); Kinder et al. (2025); Li et al. (2024); Meyer et al. (2024); Polakova and Ivenz (2024); Rashkin et al. (2025); Rüdian et al. (2025); Schaller et al. (2024); Seßler et al. (2025); Shin et al. (2025); Stahl et al. (2024); Steiss et al. (2024); Tan et al. (2024); Usher (2025); Venter et al. (2024); Wang (2024); Wang et al. (2025); Weidlich et al. (2025); Zhai and Ma (2022); Zou et al. (2025).

al., 2025; Stahl et al., 2024; Zhai & Ma, 2022), all studies examine a version of OpenAI's ChatGPT as an LLM reviewer, whereas 14 studies also use language models from other providers.

Metafeedback was frequently based on general or psychological criteria (e.g., perceived usefulness, clarity, motivation), whereas feedback derived directly from classroom practice was rare. Seven studies included real teachers as reviewers, but only one study involved an actual German language teacher.

Regarding text genres, 18 studies focused on argumentative writing; informative and instructional genres were almost entirely absent.

This brief overview already shows that the quality of and even the given feedback itself can be operationalized quite heterogeneously. Beside the fact that nearly every study that was considered explores the quality of the feedback given with statistical methods (which brings other problems), most of the studies use the comparison between human- and AI-generated feedback as their quality measure. This presents a reliability–truth dilemma of this particular field: Reliability should analyze the true quality of the feedback. But as we know, reliability should not be confused with truth, because raters can also agree on a factually incorrect opinion. And this highlights central challenges when analyzing feedback quality: How can a variable (feedback quality) be operationalized for which there is no existing ground truth?[8] How can we prevent LLMs from merely replicating and amplifying the problems that teachers also have with feedback? This leads to necessary compromises in every study.

2.4 Research gaps

One research gap that emerges is in the lack of externally valid studies conducted in Germany, particularly studies involving authentic classroom contexts, younger pupil populations, realistic writing tasks, and criteria grounded in the German curriculum. Research should also include qualified teachers as human comparators for LLMs to ensure pedagogical relevance and ecological validity.

There is also a lack of integrating the assessments of *real* teachers to *real* texts of pupils. This results in limited ecological validity of the studies reported. Therefore, the study design of Zou et al. (2025) should be emphasized: They assessed the metafeedback against the background of the students' revisions—that is, they looked specifically at what kind and which parts of the feedback were included in the revisions and how successful that was. Meyer et al. (2024) also looked at the text revisions and had the quality of this revision compared to the first text version assessed by a support vector machine specially trained for this purpose. Its agreement with the judgments of human experts showed limited validity. Of course, like other reconstructive studies, the results are also subject to certain limitations in terms of validity, particularly because Zou et al. (2025) carried out the categorizations themselves. Nevertheless, these designs are interesting for future studies, as they also consider the revisions and thus also take an extended feedback cycle into account.

Because the analysis of previous studies on feedback relates to a comparatively large number of texts, results often lack a ground truth that comes close to an objective assessment. The question is: What

[8] One problem is that not even supposedly obvious grammatical errors or spelling mistakes are as obvious as they seem (Lüdeling, 2008). Anyone who has understood this may be tempted to avoid the challenge of discussing the content of texts in the first place. Of course, this is not a target-oriented solution for didactics, but an open discourse on what we understand by an error or no error, by a more successful and less successful formulation, and finally by a good and a less good text. Of course, there are many correct answers to these questions, but we must try to answer the question regarding which of these correct answers make the most sense pedagogically and didactically.

is the actual quality of the text? This is closely related to the question of how good the feedback is. Using a significantly smaller sample size can lead to the creation of a "ground truth" or a "correct" result for the texts from a didactic point of view, despite the problems this necessarily entails in terms of informative value.

The present study will tackle the following research gaps: Specifically, there is a notable lack of studies involving didactics of German-relevant producers and products such as authentic German pupils, their texts (especially nonargumentative text forms), real teachers, real-world tasks, criteria that are used in German classes, systematic consistency analyses accounting for the probabilistic nature of AI-generated outputs, and comparative analyses across diverse AI models. Nearly all studies also work with texts in which it is unclear what the ground truth is in terms of quality. This study addresses these gaps (ecological validity for German schools and three texts with given ground truth) and expands on existing research discourse accordingly.

3 Study

The above explanations result in a multidimensional operationalization of feedback quality that often comprises the evaluation of texts in the form of points and written feedback on the texts. Although the former can be measured by a relatively simple comparison of quantitative values, the latter requires a more complex operationalization method (see 3.2). The research questions bundle these complex relationships and are therefore as follows.

3.1 Research questions

RQ1: How reliable is human- or AI-given, criteria-based scoring on pupils' texts?

RQ2: How much do humans and AI agree when it comes to the criteria-based assessment of pupils' texts?

RQ3: In which areas can differences be spotted between human- and AI-based ratings of pupils' texts?

RQ4: Which quality dimensions of feedback predict whether it is perceived as helpful by metareviewers?

3.2 Method

3.2.1 Participants/Producers

Due to data protection regulations, nothing more can be told about the three pupils (writer) who wrote the texts (source) in question other than that they attended the 6th grade of a Bavarian secondary school (*Gymnasium*), that German is their L1, and that they have not been diagnosed with any reading or spelling difficulties.

In principle, it is not easy to collect data from teachers (Reviewer 1) in Germany because they are very busy and rightly represent a group worthy of protection from too many studies. The data on human teachers were therefore collected at two points in time (April and June 2025) in similar contexts. At both points in time, the author gave a workshop on "AI in German lessons" and, after a technological introduction, every teacher assessed three texts (see next section). A total of 75 German teachers took part in the study, all of whom were actively teaching at a Bavarian secondary school (*Gymnasium*) at

the time of the study. On average, they had a teaching experience of *M* = 22 years (*SD* = 10). In addition, 70% of them worked in positions of responsibility within the school system (e.g., as German class supervisors for an entire school or as instructors for young teachers) for the State of Bavaria. This is associated with high barriers in terms of qualifications. It can therefore be assumed that this is a small but extremely competent and experienced group of German teachers that represents a unique sample in the research field.

The AI systems (Reviewer 2) for the comparison were selected to balance high-performance and yet versatile systems. This is because the section above on related work shows that OpenAI's models are dominant in these studies. In particular, the use of open models has not received too much attention in research. For this purpose, the following AI systems were prompted (see 3.2.3) and used: *Mistral Large, Llama 4 Maverick*, and *GPT-4.1*.

As metafeedback producers, eight preservice German teachers[9] who had completed more than half of their standard period of study were trained. The rater training consisted of four comprehensive sessions over the course of a month, during which participants were introduced to the study itself, the technological basics of AI systems, and psychological and German didactic research on feedback as well as task quality in relation to writing processes. The metareviewer then worked intensively with the three texts in order to get closer to a ground truth regarding the assessment criteria. Finally, on this basis, exemplary feedback texts were rated, while deviations were discussed before the actual ratings took place. In these, each rater assessed 36 sets of feedback (12 per text) using an online questionnaire (see below). The students did not know which producer had produced which feedback product.

3.2.2 Texts, task, and criteria

Three authentic German texts (source) of varying quality levels (high, medium, low) from 6th-grade pupils (genre: procedural description) were used to assess quality of feedback on results (Christophel et al., 2017). The three texts are labeled *Für, Diese*, and *Man* (see Appendix B). As Steiss et al. (2024) showed, text quality is a relevant factor for feedback quality, especially for ChatGPT 3.5, but for human reviewers as well. The relatively small number of texts has two key advantages in addition to the obvious disadvantages (see Limitations): First, this ensured that all human raters provided feedback on all texts. Second, a ground truth could be developed for the training of the metareviewers, which made it much easier for them to distinguish between adequate and inadequate feedback in terms of feedback content. These two advantages are further supported by the choice of text type and grade level, because for informative writing (as opposed to argumentative or creative writing) in lower grade levels, it is more likely to develop an objective ground truth. The three texts were given to the reviewers in random order to avoid sequence effects.

The task[10] underlying the texts was kindly provided to the researcher by an experienced German teacher in a leading position. Its authenticity strengthens the ecological validity of the results, which was the main reason this task was chosen for the present study, despite the following difficulties. Its

[9] The fact that the work experience of reviewers and metareviewers differs so greatly in this regard could not be resolved in any other way from an organizational perspective (active teachers could not be recruited for such extensive rater training). This is addressed in the limitations section.

[10] "You explain to your best friend in writing how best to bake a cookie. To do this, write a description of the process according to the rules discussed in class. Describe the process accurately and completely, using the necessary technical terms! —Pay attention to neat form and correct language! You have 60 minutes" (author's translation).

quality is questionable, which necessarily influences the quality of the texts and feedback. It only partially meets the criteria for meaningful writing tasks proposed by Bachmann and Becker-Mrotzek (2011): It has an identifiable function and a reasonably authentic occasion, and the material provided also imparts some of the necessary technical and general knowledge to the pupils. The task is reasonably clear and structured, and the text format of the target text is clear. However, there is a particular lack of integration into genuine social interaction, and there is no prospect of impact monitoring. There is also no differentiation, and the promotion of functional and pragmatic skills can be attested only partially. Differentiated role cards and actual baking with feedback from a friend or tandem partner would improve the task even further. The components of writing arrangements developed by Steinhoff (2018) are also only found here in places, which indicates again how far apart school reality and didactic expectations can be. Meeting those components is particularly relevant because, as Steinhoff (2018) points out, good feedback must be preceded by a good task. As has been shown, although the task leaves room for improvement, it is also a very authentic example of a task that is used in everyday writing lessons, and this strengthens the significance of this study for actual teaching (= ecological validity).

The criteria used for quantitative feedback were exactly the same as those imparted by the teacher in the classroom. There is also undisputable room for improvement (see below), but here too (as with the task itself), their authenticity contributed to the ecological validity of the survey.[11]

Table 1

Assessment Criteria and Explanations

Criterion	***Explanation***
Materials/Ingredients	First, list all the necessary materials/ingredients in complete sentences and without unauthorized abbreviations.
Procedure	Describe the process correctly and in the correct order of the steps.
Explanation	Explain the context to the reader clearly, and skillfully distinguish between what is important and what is unimportant.
Conclusion	In the conclusion, you give a helpful tip.
Style	Your language is factual and neutral.
Verb form	You vary between passive and constructions including the impersonal pronoun *one* when describing the process.
Word choice/expression	Your expression is accurate and you avoid word repetitions.
Tense	You choose the present tense throughout.
Sentence structure	You vary your sentences and link them with appropriate conjunctions.
Spelling	You have a good command of spelling and punctuation.

[11] Winkler (2018) explicitly calls for the scientific investigation of tasks that arise in "natural" German teaching. This article will not address the fundamental question of the purpose of German teaching research. For now, it is simply noted that natural tasks and criteria were used. These will be presented and critically evaluated below and clearly influence the results.

Structure	Your essay is clearly structured (paragraphs between introduction, main body, and conclusion).

Table 1 highlights a strong predominance of language aspects with a known imbalance that is equally evident in human and machine assessments. Some criteria combine several elements such as *Explanation*, *Expression*, *Sentence structure*, and *Spelling*, and they sometimes mix language and content, as in the case of *Materials*/*Ingredients*. These complex criteria were discussed with the teachers, and they were also reminded in the questionnaire that they should decide for themselves on the internal weighting of the criteria in these areas. Questions of relevance arise regarding the appropriateness of an objective tone when addressing a friend, the choice of tense, and the need for impersonal forms in process descriptions. Furthermore, the formulation of criteria can be imprecise—for instance, in determining what is considered "(un)important" in *Explanations*, what "helpful" means in *Conclusion*, or what qualifies as unauthorized abbreviations in *Materials*/*Ingredients*. There is no space here for further discussion. Nevertheless, a comprehensive examination of the question whether and which criteria teachers actually use in writing instructions today would be worthwhile, because it could provide important feedback for teacher professionalization.

The reviewers completed an online questionnaire in which they used the criteria to rate the texts on a 10-point Likert scale ranging from 1 (*Not met at all*) to 10 (*Met perfectly*). This high number of options can essentially be too large for humans, even though the classic 5–7-point scale (Cox, 1980) is now viewed in a more differentiated manner (Jebb et al., 2021).

The metareviewers used a set of criteria compiled from established feedback research (Hattie & Timperley, 2007; Steiss et al., 2024) to assess the quality of the feedback. In the online questionnaire, they indicated on a 4-point Likert scale ranging from 1 (*I strongly disagree*) to 4 (*I strongly agree*) the extent to which they agreed or disagreed with the following statements:

- The feedback is AI-generated.
- Errors are specifically named.
- Suggestions for improvement are made in concrete terms.
- The suggestions for improvement are sufficiently explained.
- The feedback is error-free (regarding the content).
- The feedback has more positive than negative ratings.
- The feedback is helpful.

3.2.3 Procedure

The reviewers (teachers and AI systems) were prepared for the task with almost the same prompt: to provide feedback on all three texts in the form of points for criteria, an overall grade, and a short text. One small difference is that the AI systems were also assigned the role of experienced teachers,[12] whereas the teachers were given a time limit (45 minutes for all three texts). Apart from that, the task and scenario are described similarly for humans and machines to make sure they are comparable. The prompt can be found in Appendix A and, besides the task (quantitative and qualitative feedback), it

[12] As Jacobsen and Weber (2025) show, explicit information about the context (e.g., the role of the LLM) in the prompt leads to better feedback.

includes a scenario that was as authentic as possible and designed to make it clear that the fictional pupils would use their feedback for the revision.

While each teacher provided feedback on each text only once, the AI systems provided each piece of feedback 10 times to test their consistency. For the models that are adjustable, settings (temperature, top p, and seed) were kept constant for the 10 runs. The prompts were given as user prompts (Neumann et al., 2025) and presented independently.

The data from the reviewers and the metareviewers were collected using an online questionnaire via *SosciSurvey* (Leiner, 2024), processed in *Excel*, and statistically evaluated with *R* (R Core Team, 2025).

4 Results

4.1 RQ1: How reliable is human- or AI-given criteria-based scoring on pupils' texts?

The first step of the study examines the descriptive distribution data for the teachers and the AI systems.[13] The following presentation focuses initially on the overall grades given, as these are the easiest to interpret, before going into more detail on the points for the individual criteria.

Table 2

Mean and Distribution Values of Overall Grades by AI and Teachers According to Texts

Rater	***Text***	M_{grade}	SD_{grade}	MIN_{grade}	MAX_{grade}	$Range_{grade}$
AI	*Für*	12.93	0.78	12	14	2
AI	*Diese*	10.20	1.58	8	12	4
AI	*Man*	8.67	1.42	7	12	5
Teacher	*Für*	12.67	1.29	8	15	7
Teacher	*Diese*	7.24	2.12	2	13	11
Teacher	*Man*	4.31	2.04	1	10	9

Regarding the teachers, it is initially apparent from Table 2 that they (correctly) assess three levels of the three texts (*Für* → *Diese* → *Man*), because the grades range across the different levels of quality, with them distinguishing clearly between the quality levels. At the same time, the range between the minimum and maximum values shows that there are high individual deviations (up to 11 points, which in the German grading system, which ranges from 1 to 6, corresponds to a deviation between the best and second-lowest grades) within the text *Diese*. The range of standard deviations from 1.3 up to 2 points is to be expected, given the wide range of possible points (0–15).

The AI models also divide the texts into three levels, although the overall grades are much closer to each other than the grades given by human reviewers. The low standard deviations and small ranges indicate little variance within the texts. On average, the weakest text is graded twice as highly as by the teachers.

[13] To make the results more readable, the values of the individual AI systems are summarized in some places.

To gain insights into criterion-based ratings, Table 3 shows the descriptive data for the AI systems by text and criterion. Similar to the grades, the text *Für* also performs best in terms of criteria, with standard deviations below 1 and a range below 4, indicating low dispersion. For the other two texts, too, the range is never higher than 5, and the standard deviations rarely exceed 1. Exceptions for the texts *Diese* and *Man* are the criteria of *Spelling*, *Conclusion*, and *Tense* in which the dispersion is wider than usual in both texts. The criteria *Verb form* and *Sentence structure* are rated relatively low over all texts.

Table 3

AI Criteria-Based Ratings

	Diese					***Für***					***Man***				
Criterion	***M***	***SD***	***MIN***	***MAX***	***R***	***M***	***SD***	***MIN***	***MAX***	***R***	***M***	***SD***	***MIN***	***MAX***	***R***
Materials	7.72	0.91	6	10	4	9.68	0.47	9	10	1	7.60	0.59	6	8	2
Procedure	7.58	0.68	6	9	3	9.10	0.50	8	10	2	6.68	0.57	5	8	3
Explanation	6.38	0.54	5	7	2	8.40	0.55	7	9	2	5.68	0.57	4	7	3
Conclusion	8.03	1.10	6	10	4	7.30	0.94	5	9	4	5.62	1.27	3	8	5
Style	7.12	0.72	5	8	3	8.47	0.51	8	9	1	6.22	0.73	5	8	3
Verb form	5.95	0.75	4	7	3	7.35	0.70	6	9	3	5.18	0.87	4	7	3
Expression	6.60	0.63	5	8	3	8.15	0.77	6	9	3	5.58	0.75	4	7	3
Tense	8.18	1.45	5	10	5	9.50	0.72	8	10	2	7.38	1.31	5	10	5
Sentence str.	6.10	0.63	5	7	2	8.00	0.99	6	10	4	5.35	0.58	4	7	3
Spelling	6.00	1.40	4	9	5	9.15	0.53	8	10	2	4.58	1.38	3	8	5
Structure	7.25	0.74	5	8	3	9.00	0.64	8	10	2	6.30	0.76	5	8	3

Table 4 shows the same statistical measures for teachers. Overall, the three-level classification of the texts indicated by the grades can also be understood here based on the points. At the same time, however, the range within a text is again considerably greater than in the points awarded by the AI system, both horizontally and vertically: For example, the range is rarely less than 5 and the mean values for the text *Man* for the criteria of *Spelling* (1.80) and *Tense* (8.53) are clearly apart, which could also indicate a differentiated and independent assessment of the individual criteria by the teachers. The category *Conclusion* in the text *Für* shows a particularly low mean value and a high standard deviation, which will be discussed later. The criterion *Sentence structure* is also rated relatively low over all texts.

Table 4

Teacher Criteria-Based Ratings

	Diese					*Für*					*Man*				
Criterion	***M***	***SD***	***MIN***	***MAX***	***R***	***M***	***SD***	***MIN***	***MAX***	***R***	***M***	***SD***	***MIN***	***MAX***	***R***
Materials	6.81	2.40	1	10	9	9.48	0.91	5	10	5	7.07	2.38	2	10	8
Procedure	7.32	2.01	2	10	8	9.57	0.77	6	10	4	6.20	2.32	1	10	9
Explana-tion	6.13	1.91	2	10	8	9.25	1.00	5	10	5	4.55	2.21	1	10	9
Conclusion	6.17	2.76	1	10	9	3.29	3.30	1	10	9	5.29	2.55	1	10	9
Style	7.15	2.10	2	10	8	9.09	1.71	1	10	9	6.76	2.42	1	10	9
Verb form	6.75	2.18	2	10	8	9.27	1.02	6	10	4	2.05	1.40	1	8	7
Expression	5.31	1.87	2	9	7	9.12	1.04	6	10	4	3.35	1.40	1	7	6
Tense	8.79	1.79	2	10	8	9.71	0.54	8	10	2	8.53	2.20	1	10	9
Sentence str.	4.80	2.03	1	9	8	8.91	1.28	5	10	5	2.72	1.31	1	7	6
Spelling	4.27	1.83	1	8	7	9.53	0.81	5	10	5	1.80	0.92	1	5	4
Structure	7.11	2.53	1	10	9	7.04	1.94	2	10	8	4.60	2.65	1	10	9

To examine the reliability of the grading (overall grade) in a more statistically sound manner, a two-way mixed-effects intraclass correlation coefficient for consistency, single measures (ICC (C,1)) was used (see Table 5).

Table 5

ICC Values of the Reviewers' Grades

Rater	***$N_{ratings}$***	***Texts***	***ICC***	***F test (df1, df2)***	***P***	***95% Confidence Interval (CI)***
Teacher	75	3	0.875	F(2, 148) = 528	< .001	[0.649, 0.996]
GPT-4.1	10	3	0.920	F(2, 18) = 117	< .001	[0.711, 0.998]
Mistral L	10	3	0.919	F(2, 18) = 115	< .001	[0.707, 0.998]
Llama 4	10	3	0.750	F(2, 18) = 31.1	< .001	[0.368, 0.992]

Interrater reliability among the 75 teachers who evaluated the three texts yielded an ICC of 0.875, indicating a high degree of consistency across the reviewers' grades. The associated *F* test was significant. The 95% confidence interval (CI) from the teachers [0.649, 0.996] suggests that the true level of consistency in the population is at least moderate and potentially near-perfect. These results indicate

that, although individual teachers may have different absolute scoring levels, there is strong agreement in the relative ranking of the texts.

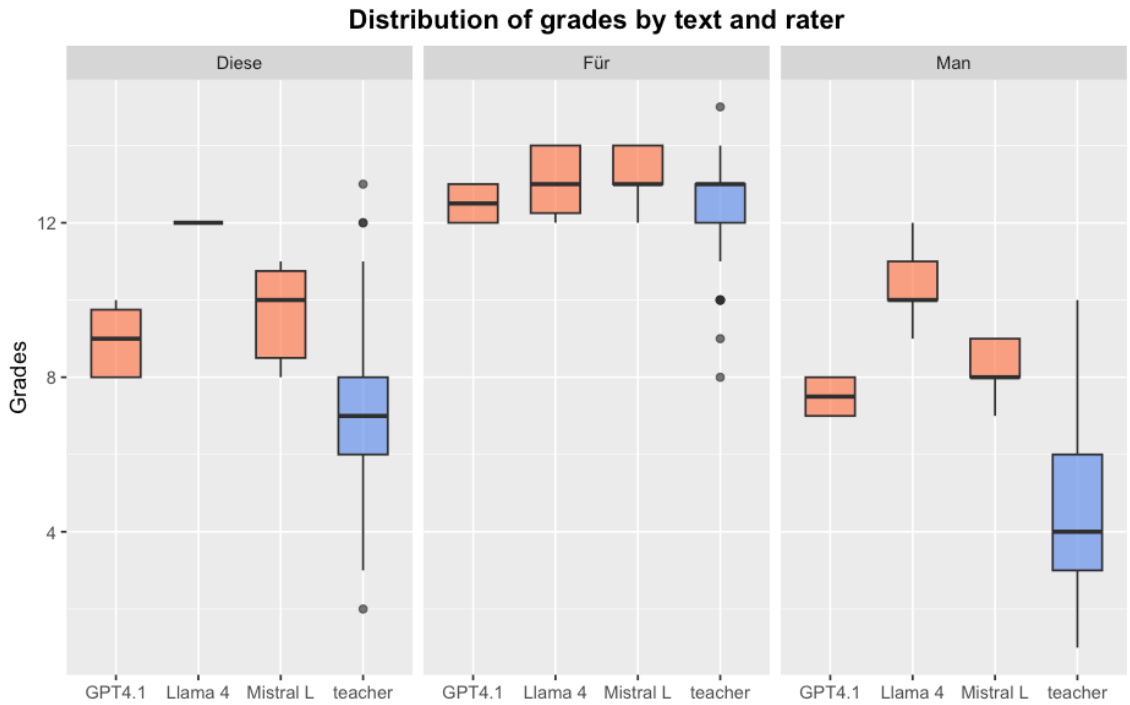
For the 10 AI ratings of the three texts, the ICC values indicate very high interrater consistency except for Llama 4. All *F* tests were significant, and the 95% CIs suggest that the true "population" consistency is at least good and may be near perfect, again except for Llama 4.

4.2 RQ2: How much do humans and AI agree when it comes to the criteria-based assessment of pupils' texts?

First, a circumstance that was already indicated by the descriptive data of the grades from RQ1 is examined graphically, namely that the teachers distinguished more clearly between the three texts. To answer RQ2, the box plots in Figure 2 show how the overall grades for the texts are distributed for each of the three texts separately according to reviewer (group).

Figure 2

Box Plots of Grades by Text and Rater



It can be noted that the medians for the AI raters are mostly closer together overall; and, for two of the three texts, they are significantly higher than for the teachers. Teachers distinguish more clearly between the three texts and are thus closer to the ground truth. For the *Für* text, the medians are roughly the same. In terms of dispersion, the height of the boxes varies, but is lowest for the best text, except for Llama4's scores for the text *Diese*, which scored 12 points 10 times and was thus significantly higher than the average scores of the other raters. The whisker lines and outlier points take into account the fact that the dispersion among teachers is generally higher for a text than among AI systems.

To gain detailed insights into the criteria-based ratings, Table 6 shows the descriptive data for the teachers' and AI systems' ratings (mean values and differences) as well as a *p* value of a Wilcoxon rank-sum test with a Holm–Bonferroni correction for multiple comparisons.

Table 6

Criteria-Based Assessments From AI and Teachers

	Diese				*Für*				*Man*			
Criterion	M_{human}	M_{AI}	***Diff***	p_{holm}	M_{human}	M_{AI}	***Diff***	p_{holm}	M_{human}	M_{AI}	***Diff***	p_{holm}
Materials	6.81	7.72	-0.91	0.85	9.48	9.68	-0.20	0.49	7.07	7.60	-0.53	0.99
Procedure	7.32	7.58	-0.25	1.00	9.57	9.10	0.47	0.00***	6.20	6.68	-0.47	0.62
Explanation	6.13	6.38	-0.24	1.00	9.25	8.40	0.85	0.00***	4.55	5.68	-1.13	0.00***
Conclusion	6.17	8.03	-1.85	0.01**	3.29	7.30	-4.01	0.00***	5.29	5.62	-0.33	0.85
Style	7.15	7.12	0.02	1.00	9.09	8.47	0.62	0.00***	6.76	6.22	0.54	0.49
Verb form	6.75	5.95	0.80	0.03*	9.27	7.35	1.92	0.00***	2.05	5.18	-3.12	0.00***
Expression	5.31	6.60	-1.29	0.00***	9.12	8.15	0.97	0.00***	3.35	5.58	-2.23	0.00***
Tense	8.79	8.18	0.61	0.03*	9.71	9.50	0.21	0.26	8.53	7.38	1.16	0.00***
Sentence str.	4.80	6.10	-1.30	0.00**	8.91	8.00	0.91	0.00***	2.72	5.35	-2.63	0.00***
Spelling	4.27	6.00	-1.73	0.00***	9.53	9.15	0.38	0.00***	1.80	4.58	-2.78	0.00***
Structure	7.11	7.25	-0.14	1.00	7.04	9.00	-1.96	0.00***	4.60	6.30	-1.70	0.00***

Overall, 22 of the 33 averaged differences show significant discrepancies between the criteria-based judgments of humans and LLMs. The judgments for the text *Für* deviate significantly from one another. The largest difference in relation to the *Conclusion* criterion is explained in the next section. The differences also vary in terms of magnitude and direction. This is because teachers assess the two weaker texts more strictly in relation to most criteria, resulting in mostly negative differences, whereas they assess the best text more positively. Although the overall boxplots for this text suggest close agreement between human and AI ratings at the global level, the criterion-level analysis reveals many statistically significant differences. This apparent paradox may arise because both groups arrive at similar total scores through different weighting of the criteria: AI systems and human raters emphasize or de-emphasize different aspects of the text, leading to compensating effects that mask criterion-specific

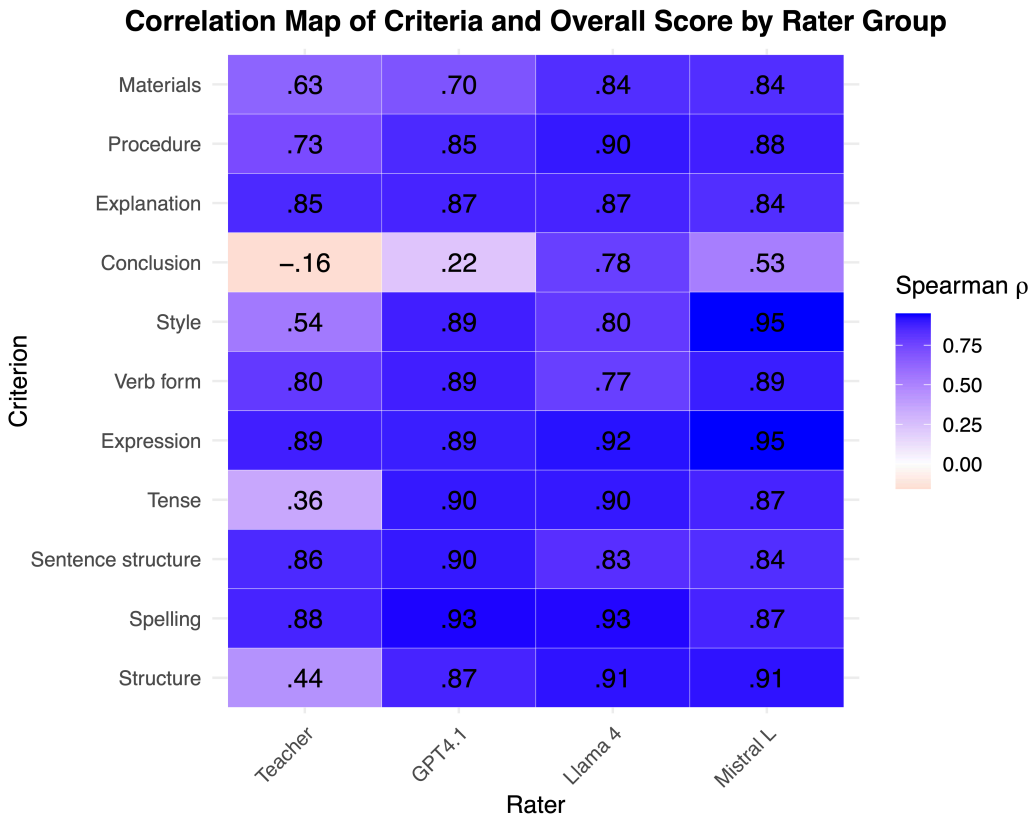
divergences in the aggregated boxplots. Although the grading tendencies of AI and teachers are similar, there are clear differences in the criteria-based assessment. These differences in grading and criteria-based assessment may be explained by the different weightings of the criteria and their influence on the overall grade. This leads directly to RQ3.

4.3 RQ3: In which areas can differences be spotted between human- and AI-based ratings of pupils' texts?

The Spearman correlation plots for teachers and AI models (see Figure 3) show how highly the individual assessment categories correlate with the overall grade:

Figure 3

Correlation Map of Criteria and Overall Score by Rater Group



Almost all values for the three AI models correlate highly with the overall score as well as with each other, although this cannot be shown in the figure. Among all raters, the *Conclusion* category correlates particularly weakly and even negatively with the overall score. Of course, this may be because the conclusion of the text is considered less important by the reviewers; but with regard to the text *Für*, another explanation suggests itself: This text, which otherwise scores very well in almost all criteria, does not contain a conclusion. This explains the highest variance values among all assessment categories in Table 4 and to the highest difference between AI and humans in Table 6.

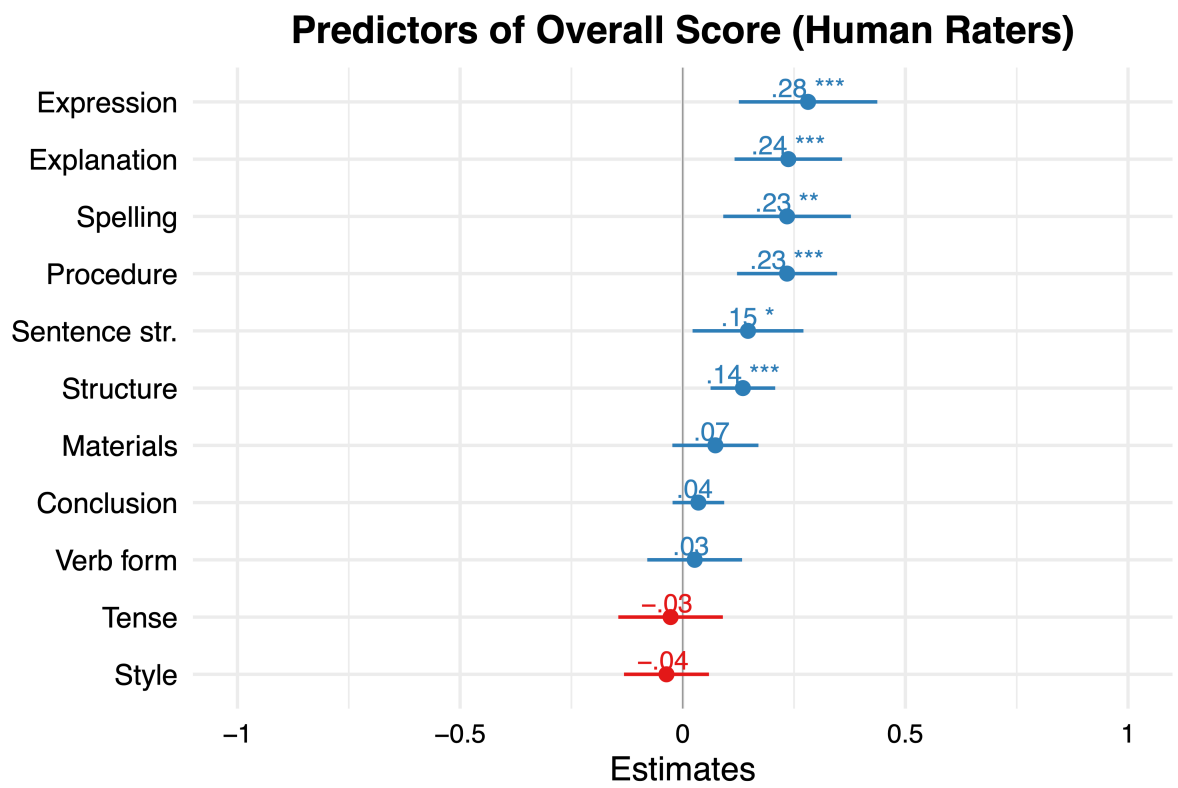
The picture is more nuanced when it comes to teachers. In addition to high scores in the language categories of *Expression*, *Sentence structure*, and *Spelling*,[14] the *Explanation* correlates particularly highly in terms of content. In contrast to the models, however, other criteria such as *Tense* or *Structure* do not show a high correlation with the overall grade among teachers.

The linear mixed-effects models employed here provide a more statistically robust analysis of RQ3. For the three AI models, these models could not be estimated meaningfully due to singular fits and perfect or near-perfect multicollinearity among predictors. Figure 3 already indicates high correlations between the rating criteria and the overall score; and also exploratory analyses revealed very high correlations ($r > .9$) between all rating criteria, resulting in overfitted models ($R^2 \approx .97$) with nonsignificant individual predictors. Because the rating criteria were not statistically independent within the AI data, no interpretable regression coefficients are reported. This indicates that the AI systems produced highly consistent, undifferentiated evaluations across criteria, effectively assigning global quality judgments rather than criterion-specific ratings.

In contrast, the teachers showed sufficient variability across criteria to allow for stable model estimation, indicating more differentiated evaluative behavior.

Figure 4

Estimates of the Fixed Effects From the Mixed-Effects Model for Human Raters



[14] The dominance of language criteria in essay assessment has already been pointed out by Birkel (2003) or more recently by Vögelin et al. (2019) and, for AI models, by Seßler et al. (2025).

Figure 4 displays the fixed effects from the linear mixed-effects model (detailed results in Appendix C) predicting overall text scores given the rating criteria. The model showed a marginal R^2 of .82 and a conditional R^2 of .90 (Nakagawa et al., 2017), indicating that the fixed effects (i.e., the rating criteria) accounted for most of the variance in overall scores, whereas random variation between raters and texts contributed modestly to the total explained variance. Positive coefficients indicate that higher scores on a criterion are associated with higher overall evaluations. Significant positive predictors included *Expression*, *Explanation*, *Spelling*, *Procedure*, *Sentence structure*, and *Structure*, whereas *Style* and *Tense* showed small, nonsignificant negative associations. All other predictors (*Materials*, *Conclusion*, *Verb form*) had negligible effects.

The advantage of this model compared to the simple correlation analysis shown in Figure 3 is that the criteria and random effects are examined in parallel in terms of their influence on the overall score. In the model, the influence of the categories is lower for *Sentence structure* and even insignificant for *Verb form*, whereas in Figure 3, they appear to have a huge influence. Another specific advantage is evident in the criterion *Procedure*, which is considerably more relevant here than in simple correlation analysis.

Random-effect variances indicated modest variability between raters (*SD* = 0.39) and somewhat greater variability between texts (*SD* = 0.87), suggesting that while raters differed slightly in their general scoring tendency, text-level differences contributed more strongly to variation in overall grades. This rests on only three texts and should not be interpreted substantively (modeling text as a fixed factor left the criterion coefficients essentially unchanged).

Human raters demonstrated differentiated weighting of content, linguistic, and structural criteria in their overall evaluations, whereas the AI systems produced highly uniform, globally consistent ratings across all criteria without meaningful distinctions.

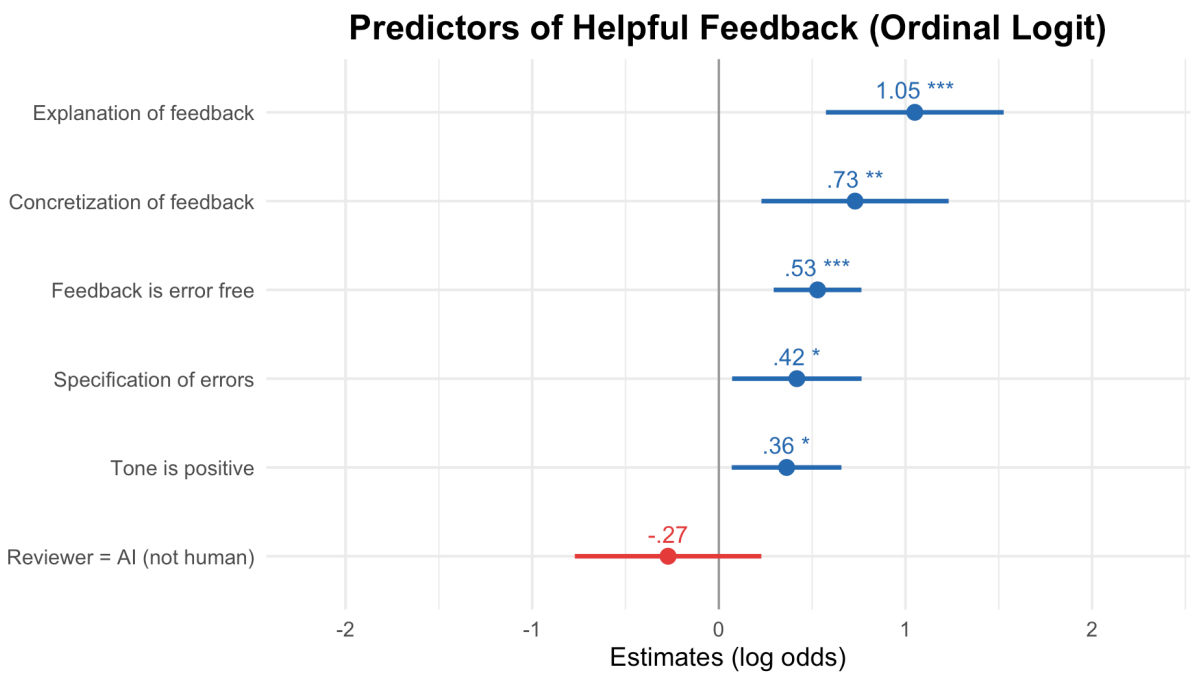
4.4 RQ4: Which quality dimensions of feedback predict whether it is perceived as helpful by metareviewers?

To examine which feedback characteristics predicted perceived helpfulness by the metareviewers, ordinal logistic regression analyses were conducted (proportional odds model, *AIC* = 536.434, $\chi^2 = 25.09$, *df* = 16, *p* = .07, $R^2 = .30$)[15] with *helpfulness* of the feedback as the dependent variable. Predictors included all other feedback criteria (see Figure 5), as well as reviewer (AI, teacher) and student text quality (high, medium, low). Odds ratios and *p* values are used to assess the relative contribution of each predictor (see Figure 5).

[15] Note that McFadden's pseudo R^2 is not directly comparable to the R^2 in other regression models (e.g., RQ3). Values between .2 and .4 are typically considered indicative of a moderate to good model fit (McFadden, 1977; Menard, 2000). The global test of the proportional odds assumption was not significant, suggesting that the assumption was met.

Figure 5

Estimated Effects (Log Odds) from the Ordinal Logit Mixed Model Predicting Helpfulness of the Feedback



The strongest predictors of perceived helpfulness by the metareviewers were the presence of sufficiently explained (*OR* ≈ 2.86, *p* < .001), concrete (*OR* ≈ 2.08, *p* = .004), and error-free feedback (*OR* ≈ 1.70, *p* < .001). Each of these dimensions substantially increased the likelihood that feedback would be rated as helpful. Additional significant predictors were the specific naming of errors (*OR* ≈ 1.52, *p* = .018) and a predominantly positive tone (*OR* ≈ 1.44, *p* = .015), both of which contributed to higher helpfulness ratings. In contrast, AI as reviewer compared to teachers as reviewers did exert a negative effect, but it is not significant (*OR* ≈ 0.76, *p* = .287). Likewise, the quality level of the underlying student text showed a nonsignificant effect for the text with the medium quality (*Diese*). However, the likelihood that the metareviewers would rate the feedback on the weak text (*Man*) as helpful is significantly higher (*OR* ≈ 2.33, *p* = .007) than for the good text (*Für*).

5 Discussion

When it comes to predicting an essay's grade, AI systems show a high level of agreement with average grades from experienced teachers. This is consistent with the results of other studies that demonstrate a moderate (Li et al., 2024; Stahl et al., 2024) up to high (Seßler et al., 2025) level of alignment between the grades assigned by humans and AI systems. Although this could lead to a recommendation in favor of preassessment, the European Union's AI Act is critical of the use of AI systems for grading (European Parliament, 2024)—and with good reason. In the debate on the use of AI systems in education, the discrepancy between legal classification and technical capability must be addressed. Additionally, the analysis of the criterion-based judgments also shows that the teachers' judgments differ significantly from the AI criterion-based ratings.

When evaluating the texts using grades or criteria-based points, it becomes apparent that teachers show significantly greater variation than the tested AI systems between different criteria, not only

within a category but also within a text. Although this may appear problematic in terms of objectivity within a category,[16] the difference between the criteria within a text in predicting the overall grade (see Figure 3) shows that teachers are able to evaluate individual criteria more independently of one another. The tested AI systems were less able to distinguish between criteria due to their iterative inferencing. As a result, AI systems used for criteria-based scoring would need even more information, and examples and would also have to assess the criteria separately from one another. To return to the *reliability–truth dilemma* mentioned at the outset: Although the reliability of AI systems (except for Llama 4) is always higher, it remains questionable whether this is also closer to the truth, because they consistently gave higher grades.

The teachers show a (surprisingly) high degree of agreement on the grades (*ICC* = 0.875), apart from a few outliers. This result broadens the scope of research into the reliability of teachers' assessments of German essays, which is still a relatively underresearched area (Birkel, 2003; Mathiebe et al., 2024; Schröter et al., 2022). The high level of agreement may be due to the combination of short, informative texts and highly experienced teachers employed in this study. Notably, although the criteria used for this study were contestable from a didactic perspective, they performed very well operationally: In the mixed model for RQ3, they accounted for the vast majority of the variance in teachers' holistic judgements (marginal R^2 = .82). In other words, these ecologically valid criteria, derived from one teacher's own practice, accurately reflected the factors that influenced the highly experienced teachers' overall grades, despite not representing optimal didactic constructs.

The cumulative link mixed model used indicates several strong predictors of perceived helpfulness by the metareviewers (e.g., Explanation, Concretization or Error-free) that are plausible and fit into the feedback model from Hattie and Timperley (2007). The fact that the model predicts that, in terms of the helpfulness of the feedback, it makes no difference to the metareviewers whether it was generated by AI or by teachers does not prove that the quality is similarly high. To ensure this, experts and the pupils themselves would also have to evaluate the feedback both before and after it was implemented in the text. Nevertheless, it can be said that the metareviewers, who recognized AI-generated feedback as such in 70% of cases, found no significant difference in terms of helpfulness. This result also seems to apply to pupils, for whom it is probably irrelevant who created the feedback message as long as it helps them. What is more relevant here is who communicates it, how they do so, and what relationship the receiving person has with the communicating person/agent. Immediacy and scalability will remain the biggest advantages of AI-generated feedback, whereas the question of the conditions for trust in the reviewer will need to be examined more closely in the future. That the feedback (no matter if AI- or human-based) is perceived as more helpful for the weakest text seems obvious, because there is much more room for improvement.

These results could help to limit the algorithm aversion bias (Jussupow et al., 2024) of teachers and pupils, which, due to exaggerated expectations, sometimes leads to the nonacceptance of supposed and actual errors made by AI systems, whereas errors are also made every day in assessments by peers or teachers. In exploratory focus group interviews, Holtzman and Nimrod (2025, p. 6) note: "Acceptance [of technology faults] is also affected by the user's expectations regarding the technology." The completely exaggerated expectations of AI systems could be adjusted to reality with the help of

[16] I believe it is far from certain that objectivity in the assessment of a pupil's text should be viewed as a purely positive construct. From a pedagogical and psychological perspective, a teacher's subjectivity can contribute to feedback being significantly more valuable than feedback that is objective and (supposedly) factually correct. This once again highlights the complexity of trying to measure the construct of feedback quality.

better education about how AI systems work. However, this should not obscure the fact that there is still a lot of work to be done in this field to improve (measuring) the quality of AI-generated feedback.

6 Limitations

Choosing only three texts for making the comparison is obviously one of the most important limitations. One reason for this limited number is that teachers (and especially those who teach prospective teachers in addition to their duties as teachers) often feel overloaded and have correspondingly little time. A compromise had to be found that allowed time-consuming narrative feedback to be collected in addition to quantitative feedback. This was not conceivable with more than three texts.

Another limitation inherent to the topic is the rapid development of AI models. Shortly after the data were collected, OpenAI introduced GPT-5. It would have been interesting not least to compare whether the improvements also apply to the area discussed here. Also, the models used were not trained especially for this task.

Another limitation is the choice of an authentic task that does not meet the contemporary understanding of what constitutes a good writing task or a writing arrangement that supports learning. This affects not only the texts, but also the feedback and metafeedback provided.

The sample of teachers was very exclusive and certainly does not represent the population as a whole, because it is a particularly experienced group. In contrast, the metareviewers are less experienced in terms of school and teaching. Operationalizing feedback quality by training student raters is not unknown in research, but that does not make it any less critical. They have not witnessed which feedback has helped which student at what level to improve their writing. This missing experience is another important limitation. Attempts were made to compensate for this through intensive rater training, and it is possible that, due to their closeness to pupils in terms of age, trainee teachers are not as unsuitable as they might appear at first glance when it comes to assessing the helpfulness of feedback. Nevertheless, it must also be noted that the selection of metareviewers is a compromise, and this leads to the most important limitation of this study.

Finally, it should also be noted that feedback quality remains a very difficult construct to operationalize, especially if the learners' perspective is not taken into account at all. The current situation, as indicated by Jansen, Horbach, and Meyer (2025) in which approximately half of the pupils do not engage with AI-generated feedback at all, should bring the perspective of feedback users much more to the fore, especially since previous studies have tended to focus on feedback-centered research (Panadero & Lipnevich, 2022). Learners have to engage with the feedback given. Otherwise, the question of how best to operationalize feedback quality remains purely theoretical.

7 Summary and Future Work

Overall, measuring the quality of AI-generated feedback and therefore operationalizing feedback quality poses many conceptual and methodological challenges. This article presents a theoretical model (Figure 1) that can help to systematize the various approaches already taken in research and tackles some of these challenges. The AI systems tested demonstrate high interrater reliability values and often agree satisfactorily with experienced teachers when it comes to the overall grade tendency for pupils' texts. Their written feedback also does not differ significantly from expert feedback provided

by the experienced teachers in terms of how helpful it is perceived to be by the trained metareviewers. However, the AI-generated grades were too lenient. Their criteria-based ratings were largely undifferentiated and diverged significantly from the ratings provided by experienced teachers.

So, is the measured AI-generated feedback good? Answering this question with a yes would overstretch the robustness of the study design and reveal an overly simplistic understanding of the processes that take place every day in schools when feedback is given. Feedback is embedded in social processes (Fong, 2025), and so an AI system will (hopefully) not know what happened in the previous PE lesson, what is currently happening at home for a child, how the pupil has developed, or what exactly was taught under what circumstances. As long as we do not provide such context data to AI systems—and I strongly advocate this—AI-generated feedback can be useful for providing teachers or students with initial guidance—for example, during exercises or by correcting orthographic errors, for example. The AI-generated feedback can therefore only be individualized and not adaptive—which is considered a key criterion for effective learning (Corno, 2008; Parsons et al., 2018; Vygotsky, 1978).

Future studies could investigate whether the advantages of AI-generated feedback, such as immediacy and scalability, could compensate for some of these gaps. Additionally, they could examine whether hybrid approaches that combine the strengths of both AI-generated and human feedback represent useful alternatives. Ultimately, AI-driven feedback is here to stay, so measuring its quality will be a demanding but worthwhile task for science, as it should be as effective as possible in supporting learning.

References

Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., & Almegren, R. M. (2024). Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International*, 1–16. https://doi.org/10.1080/14703297.2024.2437122

Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. *Journal of Educational Computing Research, 63*(2), 431–463. https://doi.org/10.1177/07356331241307297

Andringa, S., & Godfroid, A. (2020). Sampling Bias and the Problem of Generalizability in Applied Linguistics. *Annual Review of Applied Linguistics, 40*, 134–142. https://doi.org/10.1017/S0267190520000033

Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *American Psychologist, 63*(7), 602–614. https://doi.org/10.1037/0003-066x.63.7.602

Aydın, B., Kışla, T., Elmas, N. T., & Bulut, O. (2025). Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System, 133*, 103784. https://doi.org/10.1016/j.system.2025.103784

Bachmann, T., & Becker-Mrotzek, M. (2011). Schreibaufgaben situieren und profilieren. In T. Pohl & T. Steinhoff (Eds.), *Textformen als Lernformen* (pp. 191–209). Gilles & Francke.

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*(1), Article 23.

Bannò, S., Vydana, H. K., Knill, K. M., & Gales, M. J. (2024). Can GPT-4 do L2 analytic assessment? *arXiv*. https://arxiv.org/abs/2404.18557

Birkel, P. (2003). Aufsatzbeurteilung – ein altes Problem neu untersucht. *Didaktik Deutsch, 8*(15), 46–63. https://didaktik-deutsch.de/index.php/dideu/article/view/189

Bui, N. M., & Barrot, J. (2025). Using generative artificial intelligence as an automated essay scoring tool: A comparative study. *Innovation in Language Learning and Teaching*, 1–16. https://doi.org/10.1080/17501229.2025.2521003

Chiang, C.-H., & Lee, H.-Y. (2023). Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. (Volume 1: Long Papers)* (pp. 15607–15631). Association for Computational Linguistics.

Christophel, E., Baadte, C., Heyne, N., & Schnotz, W. (2017). Diagnostische und didaktische Kompetenz von Lehrkräften zur Förderung der Text-/Bild-Integrationsfähigkeit bei Schülerinnen und Schülern der Sekundarstufe I (DIKOL). In C. Gräsel & K. Trempler (Eds.), *Entwicklung von Professionalität pädagogischen Personals: Interdisziplinäre Betrachtungen, Befunde und Perspektiven* (pp. 263–281). Springer. https://doi.org/10.1007/978-3-658-07274-2_14

Corno, L. (2008). On teaching adaptively. *Educational Psychologist, 43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*(4), 407–422. https://doi.org/10.1177/002224378001700401

Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers & Education: Artificial Intelligence, 7*, 100299.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(1), Article 57. https://doi.org/10.1186/s41239-023-00425-2

European Parliament. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).* http://data.europa.eu/eli/reg/2024/1689/oj

Fix, M. (2004). Textfeedback in der Sekundarstufe I. In G. Bräuer (Ed.), *Schreiben(d) lernen: Ideen und Projekte für die Schule* (pp. 120–132). edition Körber-Stiftung.

Fokides, E., & Peristeraki, E. (2025). Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek. *Education and Information Technologies, 30*(2), 2577–2621. https://doi.org/10.1007/s10639-024-12912-8

Fong, C. J. (2025). A renaissance in feedback science? Reviewing and reimagining feedback research methods. *Contemporary Educational Psychology, 83*, 102414. https://doi.org/10.1016/j.cedpsych.2025.102414

Fürstenberg, M., & Müller, H.-G. (2024). KI im Deutschunterricht. *Der Deutschunterricht*, *2024*(5), 2–13.

Graham, S., Harris, K. R., Fidalgo Redondo, R., Harris, K., & Braaksma, M. (2017). Evidence-based writing practices: A meta-analysis of existing meta-analyses writing. In R. F. Redondo, K, Harris, & M. Braaksma (Eds.), *Design principles for teaching effective* (pp. 13–37). Brill. https://doi.org/10.1163/9789004270480_003

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education, 44*(8), 1237–1252. https://doi.org/10.1080/02602938.2019.1599815

Holtzman, I., & Nimrod, G. (2025). Forgiveness in human-machine interaction. *Frontiers in Computer Science, 7*. https://doi.org/10.3389/fcomp.2025.1617471

Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J. (2024). Comparing generative AI and expert feedback to students' writing: Insights from student teachers. *Psychologie in Erziehung und Unterricht, 71*(2), 80–92.

Jansen, T., Höft, L., Bahr, J. L., Kuklick, L., & Meyer, J. (2025). Constructive feedback can function as a reward: Students' emotional profiles in reaction to feedback perception mediate associations with task interest. *Learning and Instruction, 95*, 102030. https://doi.org/10.1016/j.learninstruc.2024.102030

Jansen, T., Horbach, A., & Meyer, J. (2025). Feedback from generative AI: Correlates of student engagement in text revision from 655 classes from primary and secondary school. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 831–836). https://doi.org/10.1145/3706468.3706494

Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.637547

Jung, J., Lu, M., Benker, S. C., & Darici, D. (2025). How model size, temperature, and prompt style affect LLM-human assessment score alignment. *arXiv*. https://arxiv.org/abs/2509.19329

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly, 48*(4), 1575–1590. https://doi.org/10.25300/misq/2024/18512

Kaliisa, R., Misiejuk, K., López-Pernas, S., & Saqr, M. (2025). How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions. *Educational Psychology*, 1–32. https://doi.org/10.1080/01443410.2025.2553639

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intra-rater reliability, and alignment with human scores. *TESOL Quarterly, 59*. https://doi.org/10.1002/tesq.3405

Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers & Education: Artificial Intelligence, 8*, 100349. https://doi.org/10.1016/j.caeai.2024.100349

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In Association for Computing Machinery (Ed.),

CHI'25: *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–23). ACM. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Lehnen, K. (2023). Peerfeedback beim schulischen Schreiben: Grundlagen, Methoden, Modellierung. *IDE: Informationen zur Deutschdidaktik, 47*(2), 18–30.

Leiner, D. (2024). *SoSci Survey* (Version 3.5.02). https://www.soscisurvey.de

Li, T., Liu, Z., Matsumura, L., Wang, E., Litman, D., & Correnti, R. (2024). Using large language models to assess young students' writing revisions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 365–380). Association for Computational Linguistics.

Limpert, A.-L. (2025, June 26). fobizz, schulKI und Co: Welche KI-Tools können Schulen nutzen? *Deutsches Schulportal*. https://deutsches-schulportal.de/unterricht/fobizz-schulki-und-co-welche-ki-tools-koennen-schulen-nutzen/

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In M. Walter & P. Grommes (Eds.), *Fortgeschrittene Lernervarietäten* (pp. 119–140). Niemeyer.

Mathiebe, M., Grabowski, J., & Becker-Mrotzek, M. (2024). Reliabilität bei der Beurteilung von Textqualität. In I. Petersen, R. Reble, & J. Kilian (Eds.), *Texte schreiben in allen Unterrichtsfächern. Textbeurteilung als Grundlage für Schreibförderung und Leistungsbewertung* (pp. 253–276). Waxmann.

McFadden, D. (1977). Quantitative methods for analysing travel behaviour of individuals: Some recent developments. *Cowles Foundation Discussion Paper* (No. 474), 1–47. https://elischolar.library.yale.edu/cowles-discussion-paper-series/707

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17–24. https://doi.org/10.2307/2685605

Meyer, J., Jansen, T., Daumiller, M., & Fleckenstein, J. (2025). Understanding individual differences in students' responses to technology-based feedback on a writing task: The role of achievement motives and initial task performance. *Journal of Research on Technology in Education*, 1–31. https://doi.org/10.1080/15391523.2025.2471765

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers & Education: Artificial Intelligence, 6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Müller, H.-G., & Fürstenberg, M. (2023). Der Sprachgebrauchsautomat: Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. *Mitteilungen des Deutschen Germanistenverbandes, 70*(4), 327–345.

Müller, N., Utesch, T., & Busse, V. (2023). Qualität statt Quantität? Zum Zusammenhang von Schreibförderungs- und Feedbackpraktiken mit Textqualität unter Berücksichtigung von migrationsbedingter Mehrsprachigkeit. *Unterrichtswissenschaft, 51*(2), 169–198. https://doi.org/10.1007/s42010-023-00173-2

Mußmann, F., Hardwig, T., Riethmüller, M., Klötzer, S., & Peters, S. (2020). *Arbeitszeit und Arbeitsbelastung von Lehrkräften an Frankfurter Schulen 2020: Endbericht*. https://doi.org/10.3249/ugoe-publ-7

Mußmann, F., Riethmüller, M., Hardwig, T., Peters, S., Parciak, M., Ohms, I. C., & Klötzer, S. (2016). *Niedersächsische Arbeitszeitstudie: Lehrkräfte an öffentlichen Schulen 2015/2016: Ergebnisbericht*. http://resolver.sub.uni-goettingen.de/purl/?webdoc-3971

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Neumann, A., Kirsten, E., Zafar, M. B., & Singh, J. (2025). Position is power: System prompts as a mechanism of bias in large language models (LLMs). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 573–598). https://doi.org/10.1145/3715275.3732038

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review, 35*, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., Pierczynski, M., & Allen, M. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research, 88*(2), 205–242. https://doi.org/10.3102/0034654317743198

Philipp, M. (2023). Formatives Feedback aus der Sicht des selbstregulierten Lernens: Grundlagen und Grundsätze förderlicher Rückmeldungen. *IDE: Informationen zur Deutschdidaktik, 47*(2).

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education, 11*(1), 2410101. https://doi.org/10.1080/2331186X.2024.2410101

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.2). R Foundation for Statistical Computing. https://www.R-project.org/

Rashkin, H., Clark, E., Huot, F., & Lapata, M. (2025, July). Help me write a story: Evaluating LLMs' ability to generate writing feedback. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 25827–25847.

Reichardt, A., Kruse, N., & Lipowsky, F. (2022). Textüberarbeitung mit Schreibkonferenz oder Textlupe: Zum Einfluss der Schreibumgebung auf die Qualität von Schülertexten. *Didaktik Deutsch*, *36*. https://didaktik-deutsch.de/index.php/dideu/article/view/449

Rensfeldt, A. B., & Rahm, L. (2023). Automating teacher work? A history of the politics of automation and artificial intelligence in education. *Postdigital Science and Education, 5*(1), 25–43. https://doi.org/10.1007/s42438-022-00344-x

Rüdian, S., Podelo, J., Kužílek, J., & Pinkwart, N. (2025). Feedback on feedback: Students' perceptions of feedback from teachers and few-shot LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*.

Schaller, N.-J., Ding, Y., Horbach, A., Meyer, J., & Jansen, T. (2024). Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 210–221). Association for Computational Linguistics.

Schröter, P., Söldner, H., Hoffmann, L., Riemenschneider, A., Jost, J., & Wieser, D. (2022). Wie vergleichbar sind die Bewertungen von Abiturarbeiten im Fach Deutsch? Empirische Studien zu verschiedenen Bewertungsmodellen. In L. Hoffmann, P. Schröter, A. Groß, S. M. Schmid-Kühn, & P.

Stanat (Eds.), *Das unvergleichliche Abitur. Entwicklungen – Herausforderungen – Empirische Analysen* (pp. 213–250). wbv.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2022). *Lehrkräfteeinstellungsbedarf und -angebot in der Bundesrepublik Deutschland 2021–2035: Zusammengefasste Modellrechnungen der Länder*. https://www.kmk.org

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In Association for Computing Machinery (Ed.), *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 462–472). Association for Computing Machinery.

Shin, I., Hwang, S. B., Yoo, Y. J., Bae, S., & Kim, R. Y. (2025). Comparing student preferences for AI-generated and peer-generated feedback in AI-driven formative peer assessment. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*.

Skar, G. B., Graham, S., & Rijlaarsdam, G. (2022). Formative writing assessment for change – Introduction to the special issue. *Assessment in Education: Principles, Policy & Practice, 29*(2), 121–126. https://doi.org/10.1080/0969594X.2022.2089488

Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*.

Steinhoff, T. (2018). Schreibarrangements: Impulse für einen lernförderlichen Schreibunterricht. *Der Deutschunterricht*, (3), 2–11.

Steinhoff, T. (2025). Künstliche Intelligenz als Ghostwriter, Writing Tutor und Writing Partner. Zur Modellierung und Förderung von Schreibkompetenzen im Zeichen der Automatisierung und Hybridisierung der Kommunikation am Beispiel des Schreibens mit ChatGPT in der 8. Klasse. In C. Albrecht, J. Brüggemann, T. Kretschmann, & C. Meier (Eds.), *Personale und funktionale Bildung im Deutschunterricht. Theoretische, empirische und praxisbezogene Perspektiven* (pp. 85–99). Metzler. https://doi.org/10.1007/978-3-662-69640-8

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894.

Sturm, A. (2016). Beurteilen und Kommentieren von Texten als fachdidaktisches Wissen. *lesefo-rum.ch*, (3), 115–132. https://doi.org/10.26041/fhnw-1001

Tan, M., Mah, C., & Demszky, D. (2024). Reframing authority: A computational measure of power-affirming feedback on student writing. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*.

Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: A comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*, 50(6), 912–927.

Venter, J., Coetzee, S. A., & Schmulian, A. (2024). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment & Evaluation in Higher Education*, 50(4), 516–536.

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing, 39*, 50–63. https://doi.org/10.1016/j.asw.2018.12.003

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Wang, D. (2024). Teacher- versus AI-generated (Poe application) corrective feedback and language learners' writing anxiety, complexity, fluency, and accuracy. *The International Review of Research in Open and Distributed Learning, 25*(3), 37–56. https://doi.org/10.19173/irrodl.v25i3.7646

Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y., & Wang, R. (2025). Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. *Computers & Education: Artificial Intelligence, 9*, 100481. https://doi.org/10.1016/j.caeai.2025.100481

Weidlich, J., Gotsch, F., Schudel, K., Marusic-Würscher, C., Mazzarella, J., Bolten, H., Bütler, D., Luger, S., Wohlfender, B., & Merki, K. M. (2025). Teacher, peer, or AI? Comparing effects of feedback sources in higher education. *Computers & Education Open, 9*, 100300. https://doi.org/10.1016/j.caeo.2025.100300

Winkler, I. (2018). Aufgaben. In J. M. Boelmann & L. König (Eds.), *Empirische Forschung in der Deutschdidaktik* (pp. 39–52). Schneider Verlag Hohengehren.

Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning, 35*(9), 2817–2842.

Zou, S., Kai, G., Jun, W., & Liu, Y. (2025). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. *Computer Assisted Language Learning*, 1–30. https://doi.org/10.1080/09588221.2024.2447279

Author's address:

Maurice Fürstenberg, Ludwig-Maximilians-Universität München, Schellingstraße 3, 80799 München, Germany
m.fuerstenberg@lmu.de

Appendix A

Sie sind Deutschlehrer mit 21 Jahren Berufserfahrung. Es ist Montag, der 23. März 2026. Kurz nach den in einer Woche beginnenden Osterferien schreibt Ihre Klasse 6b, die durchschnittliche Leistungen zeigt, die dritte Schulaufgabe (Klassenarbeit) in diesem Jahr, zum materialgestützten Informieren in Form einer Vorgangsbeschreibung. Die Schülerinnen und Schüler haben zur Vorbereitung eine Backanleitung als Übungsaufsatz geschrieben und hatten dafür eine Backanleitung als Bild zur Vorlage. Diese Aufsätze sollen die Schülerinnen und Schüler nach Ihrem Lehrer-Feedback noch einmal überarbeiten und erneut einreichen, damit Sie die Aufsätze in den Osterferien in Ruhe korrigieren können.
Da Sie die Klasse wegen dringender Dienstaufgaben diese Woche nicht mehr sehen, haben Sie sich dazu entschlossen, Ihnen eine erste Rückmeldung zu ihren Texten zu geben, und zwar in Form von …

- … Punkten zu Bewertungskriterien, welche die Klasse aus dem Unterricht kennt,
- … einer Gesamtnote sowie
- … einem Feedbacktext.

Auf dieser Grundlage sollen die Schülerinnen und Schüler ihre Übungsaufsätze überarbeiten und Ihnen bis Ende der Woche ins Fach legen.
Sie haben bereits fast alle Übungsaufsätze fertig, nur drei fehlen Ihnen noch. Sie haben aber nur noch wenig Zeit, weil Sie dann zu einem dringenden Arzttermin aufbrechen müssen. Der Rest der Woche ist mit anderen Aufgaben schon voll. Sie müssen das Feedback für die Texte also in den nächsten Minuten fertigstellen.

Ich gebe Ihnen im Folgenden die Aufgabenstellung, die Beurteilungskriterien und die Beschreibung des Bildes der Backanleitung, die die Schüler bekommen haben.

Aufgabenstellung:
Du erklärst deinem besten Freund/deiner besten Freundin schriftlich, wie man am besten einen Cookie zubereitet.
Fertige dazu nach den im Unterricht besprochenen Regeln eine Vorgangsbeschreibung an. Stelle den Vorgang genau und vollständig dar und verwende die nötigen Fachbegriffe! – Achte auf saubere Form und Sprachrichtigkeit! Du hast 60 Minuten.

Die Beurteilungskriterien lauten wie folgt:

- Material/Zutat: Du nennst zunächst alle benötigen Materialien/Zutaten in ganzen Sätzen und ohne unerlaubte Abkürzungen
- Ablauf: Den Vorgang gibst du korrekt und in der richtigen Reihenfolge der Handlungsschritte wider.
- Erklärung: Dabei erklärst du dem Leser die inhaltlichen Zusammenhänge gut und unterscheidest gekonnt zwischen Wichtigem und Unwichtigem.
- Schluss: Im Schluss gibst du einen hilfreichen Tipp.
- Stil: Deine sprachliche Darstellung ist sachlich und nüchtern.
- Verbform: Du variierst bei der Vorgangsbeschreibung zwischen Passiv- und man-Konstruktionen.
- Wortwahl/Ausdruck: Dein Ausdruck ist treffend und du vermeidest Wortwiederholungen.
- Tempus: Die gewählte Zeitform ist durchgehend das Präsens.
- Satzbau: Du gestaltest die Sätze abwechslungsreich und verknüpfst sie mit passenden Subjunktionen.
- Orthografie: Rechtschreibung und Zeichensetzung beherrschst du sicher.
- Aufbau: Dein Aufsatz ist klar gegliedert (Absätze zwischen Einleitung, Hauptteil und Schluss).

Die Backanleitung als Bild, welche die Schüler bekommen haben, enthielt Folgendes:
Oben links befindet sich der große, zweifarbige Titel: „CHOCOLATE CHIP COOKIES“
Darunter liegt ein blauer Teller mit einem Stapel Zeichentrick-Cookies; kleine Striche deute Duft oder Glanz an.
Es folgt die Überschrift „DU BRAUCHST“ in braun-roter Schrift, rechts daneben der handgeschriebene Zusatz „für ca. 20 Stück“.
Abgebildete Zutaten (von links nach rechts)

- Sack „MEHL“ und Messbecher „250 g“
- Schachtel „BACKPULVER“ mit Löffel „1 TL“
- Text „SALZ“, Salzstreuer und zwei Löffel „2 TL“
- Text „BUTTER“, Messbecher „250 g“, Butterstück, kleiner Hinweis („GESCHMOLZEN, ABER NICHT HEISS“)
- Flasche „VANILLE“ mit Löffel „1 EL“
- Schüssel „SCHOKOSTÜCKCHEN!“ „350 g“
- Glas „ZUCKER“ mit Messbecher „150 g“
- Pappschachtel „BRAUNER ZUCKER“ mit Messbecher „150 g“
- Zwei Eier mit Beschriftung „2 EIER“
- Beutel „KOKOS-RASPELN“ mit Schälchen „80 g“

Rechte Bildhälfte – Schritt-für-Schritt-Illustration

1. Gebogener Pfeil und Wort „OFEN“ zeigen auf einen Drehknopf; darüber steht „AUF 180°“.
2. Rote Schüssel; hinein fließen „BRAUNER ZUCKER“, „BUTTER“ und „ZUCKER“. Unter der Schüssel: „VERRÜHREN!“; daneben ein Schneebesen.
3. Gelber Pfeil zu einer zweiten roten Schüssel, in der ein Schneebesen liegt. Oben ein Löffel „VANILLE“ und ein aufgeschlagenes Ei sowie ein zweites Ei; Text auf der Schüssel: „UNTER RÜHREN NACH UND NACH VANILLE UND EIER DAZUGEBEN“.
4. Pfeil zu blauer Schüssel mit Aufschrift „IN EINER ANDEREN SCHÜSSEL“; hinein fallen „MEHL“ und „BACKPULVER“. Daneben steht „1 TL SALZ“ und unter der Schüssel unter einem Teigschaber: „GUT VERRÜHREN!“.
5. Inhalt einer blauen Schüssel mit senkrechtem Wort „TROCKEN“ kippt in rote Schüssel mit Wort „FEUCHT“, in der eine Masse und ein Schneebesen sind; daneben der Text „LANGSAM UNTERRÜHREN“.
6. Päckchen „KOKOS-RASPELN“ und ein gelber Becher mit Schokostückchen werden in eine rote Schüssel geschüttet; daneben ein Teigschaber.
7. Teighäufchen werden aus einer roten Schüssel mit einem Löffel auf ein „UNGEFETTETES BACKBLECH“ gesetzt; unter dem Blech steht handschriftlich „Abstand!“.
8. Eine Hand streut Salz auf ein Teighäufchen; schräg daneben: „AUF JEDEN KEKS EINE WINZIGE PRISE SALZ“.
9. Küchenuhr mit Beschriftung „ETWA 10 MIN.“ (Ziffer 10 auf der Skala). Darunter ein fertig gebackener Cookie in gelbem Strahlenkranz mit Text „BIS SIE perfekt SIND!“.
10. Ganz unten mittig: umkreiste „10“ und handschriftlicher Hinweis „abkühlen lassen!“.

Ihre Aufgabe ist nun, die Schülertexte, die per Prompt eingegeben werden in einer Tabelle nach den Kriterien zwischen 0–10 Punkten zu bepunkten, danach eine Gesamtnote zwischen 0–15 Punkten zu verteilen und einen knappen Feedbacktext zu schreiben. Geben Sie nur diese 3 Dinge (Tabelle mit Punkten (2. Spalte) nach Kriterien (1. Spalte), Gesamtnote und kurzer Feedbacktext) aus.

[text]

Appendix B

Diese Zutaten werden benötigt für 20 Kekse 250 g Mehl, ein Teelöffel Backpulver, zwei Teelöffel Salz, 250 g Butter (sie muss geschmolzen, aber nicht heiß sein), 350 g Schokostückchen, 150 g Zucker, 150 g Brauner Zucker, zwei Eier, 80 g Kokosraspeln und ein Esslöffel Vanilleextrakt.
Diese Küchengeräte werden benötigt ein Ofen, zwei große Schüsseln zwei kleine Schüsseln, ein Schneebesen, ein Schabber und ein Backblech.
Zuerst muss man den Ofen auf 180 Grad vorheizen. Dann wird in einer Schüssel der braune Zucker, die geschmolzene Butter und der Zucker mit einem Schneebesen verrührt. Danach wird unter standigem rühren zuerst der Teelöffel Vanilleextrakt und dann die zwei Eier dazugegeben. In einer anderen Schüssel wird ein teelöffel Salz, ein Teelöffel Backpulver und Mehl mit einem Schabber gut verrührt. Und trocknen lassen. Dann gibt man es in die andere Schüssel wo die geschmolzene Butter, der braune Zucker und der normale Zucker vermengt mit den Eiern und dem Vanilleextrakt sind. Das tut man unter standigem Unterrühren. Nun werden die Kokosraspeln und die Schokostückchen hinzu gegeben. Dan formt man sie zu kleinen Kügelchen und legt sie auf ein ungefettes Backblech. Streue auf jedes Kügelchen noch eine prise Salz. Als nächstes kommen sie für etwa 10 Minuten in den Ofen. Danach nur noch abkühlen lassen und genießen.
Hier noch ein Tipp: Lege die Kügelchen in einem Abstand von fünf cm auf das Backblech.

Man benötigt für 20 Kekse 250 Gramm Mehl, ein Teelöffel Backpulver, zwei Teelöffel Salz, 250 Gramm Butter (sie muss geschmolzen, aber nicht heiß sein), 350 Gramm Schokostückchen, 150 Gramm Zucker, 150 Gramm Brauner Zucker, zwei Eier, 80 Gramm Kokosraspeln, einen Esslöffel Vanille, einen Ofen, einen Schneebesen, einen Teigschaber und zwei Schüsseln.
Als erstes heizt man den ofen auf 180 Grad Oberunterhize vor.
Als Nähste gibt man den Braunen Zucker, die Butter und den Weißen Zcker in eine der beiden schüsseln und verüht alles gut.
Zunächst Rührt man noch die Vanille und die Eier unter.
In die andere Schüssel gibt man mehl, einen Teeloffel salz und Backpulver und verurt dies auch gut.
Als fünftes rührt man das trockene aus der Zweiten schüssel in die erste.
Darauf hin giebt man die Raspeln und die Schokostücke in den Teig.
Nun muss man nur noch den teig auf die ungefettete bleche verteilen und jeden Chocolate chip Cookie mit einer priese salz bestreuen, in den ofen geben und Zehn minuten warten und sie abkühlen lassen.App
hir noch ein paar tips:
Tipp ein: bestreue die cookies auch noch wenn sie heiß sind mit ein wenig Zucker.
Tipp zwei: das rezept ist fur 20 Choclate chip cookies.

Für 20 Chocolate Chips Cookies werden 250 Gramm Mehl, ein Teelöffel Backpulver, zwei Teelöffel Salz, 250 Gramm Butter, die geschmolzen, und abgekühlt wurde, 350 Gramm Schokostückchen, 150 Gramm Zucker, 150 Gramm Brauner Zucker, zwei Eier, 80 Gramm Kokosraspeln und ein Esslöffel Vanilleextrakt benötigt.
Als Materialien werden zwei Rührschüsseln, ein Teelöffel, ein Esslöffel, ein Schneebesen, ein Teigschaber und ein ungefettetes Backblech verwendet.
Bevor man anfängt zu backen, wird der Ofen auf 180 Grad vorgeheizt. Zuerst werden weißer Zucker, brauner Zucker und Butter in die Rührschüssel geschüttet und mit Hilfe eines Schneebesens miteinander vermischt. Unter ständigen Rühren werden auch die Eier und das Vanilleextrakt in die Masse gegeben und so vermengt, dass der Teig zu einer homogenen Masse wird. In die andere Schüssel füllt man nun das Mehl und verrührt es mit einem Teelöffel Salz und dem Backpulver mit Hilfe eines Teigschabers. Nun wird die Mischung der trockenen Zutaten langsam in die feuchte Masse untergehoben. Jetzt werden auch die Kokosraspeln und die Schokoladenstückchen hineingerieselt und mit dem Teigschaber in der Masse untergerührt, sodass die Stücke gleichmäßig verteilt sind. Der Teig wird nun in kleinen Portionen auf das Backblech verteilt. Hierfür sollte man einen Teelöffel verwenden, da ausreichend Abstand zwischen den Teigkügelchen verhindern kann, dass diese während des Backvorgangs zusammenkleben. Vor dem Backen wird auf jeden noch ungebackenen Keks eine winzige Prise Salz gestreut. Zum Schluss schiebt man die Bleche mit den Keksen in den Ofen. Sie werden für etwa zehn Minuten gebacken - und anschließend abgekühlt. Jetzt kann man die Chocolate Chip Cookies genießen.

Appendix C

Table C1

Random Effects

Group	***Variance***	***SD***
Rater	0.16	0.38
Text	0.75	0.86

Fixed Effects (Human Raters)

Term	***Estimate***	***SE***	***df***	***t***	***p***
(Intercept)	-0.22	0.76	5.75	-0.29	.78
Materials	0.07	0.05	211.70	1.49	.14
Procedure	0.23	0.06	210.44	4.11	.00
Explanation	0.24	0.06	211.59	3.87	.00
Conclusion	0.04	0.03	212.84	1.19	.23
Style	-0.04	0.05	209.68	-0.76	.45
Verb form	0.03	0.05	193.08	0.50	.62
Expression	0.28	0.08	203.11	3.57	.00
Tense	-0.03	0.06	200.67	-0.46	.65
Sentence str.	0.15	0.06	209.66	2.32	.02
Spelling	0.23	0.07	161.14	3.22	.00
Structure	0.14	0.04	211.47	3.67	.00

Table C2

Random Effects (Standard Deviations)

Group	*Component*	*Value*
ID	*SD* (Intercept)	0.39
Text	*SD* (Intercept)	0.87
Residual	*sd*__Observation	1.11

Fixed Effects

Term	*Estimate*	*OR*	*SE*	*p*
Explanation of feed-back	1.05	2.86	0.243	<0.001
Concretization of feed-back	0.730	2.08	0.256	0.004
Feedback is error-free	0.529	1.70	0.120	<0.001
Specification of errors	0.418	1.52	0.177	0.018
Tone is positive	0.363	1.44	0.150	0.015
Reviewer = AI	-0.272	0.762	0.255	0.287
Text quality low	0.845	2.33	0.314	0.007
Text quality medium	0.244	1.28	0.305	0.423