Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring

Kathrin Seßler
Technical University of Munich
Munich, Germany
kathrin.sessler@tum.de

Maurice Fürstenberg
University of Tübingen
Tübingen, Germany
maurice.fuerstenberg@uni-tuebingen.de

Babette Bühler
Technical University of Munich
Munich, Germany
babette.buehler@tum.de

Enkelejda Kasneci
Technical University of Munich
Munich, Germany
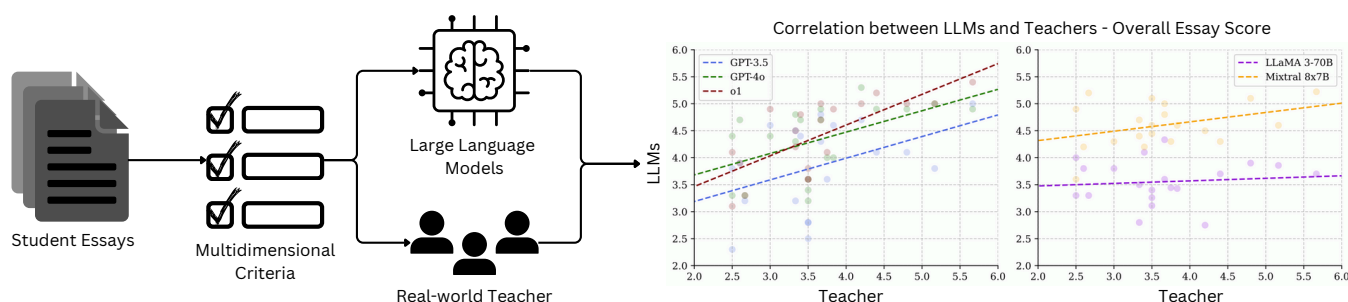enkelejda.kasneci@tum.de

Figure 1: Design and workflow of our study: student essays are evaluated based on predefined, multidimensional criteria by both Large Language Models and human teachers, followed by a comprehensive analysis of the resulting ratings.

Abstract

The manual assessment and grading of student writing is a time-consuming yet critical task for teachers. Recent developments in generative AI offer potential solutions to facilitate essay-scoring tasks for teachers. In our study, we evaluate the performance (e.g. alignment and reliability) of both open-source and closed-source LLMs in assessing German student essays, comparing their evaluations to those of 37 teachers across 10 pre-defined criteria (i.e., plot logic, expression). A corpus of 20 real-world essays from Year 7 and 8 students was analyzed using five LLMs: GPT-3.5, GPT-4, o1-preview, LLaMA 3-70B, and Mixtral 8x7B, aiming to provide in-depth insights into LLMs' scoring capabilities. Closed-source GPT models outperform open-source models in both internal consistency and alignment with human ratings, particularly excelling in language-related criteria. The o1 model outperforms all other LLMs, achieving Spearman's $r = .74$ with human assessments in the *Overall* score, and an internal consistency of $ICC = .80$, though biased towards higher scores. These findings indicate that LLM-based assessment can be a useful tool to reduce teacher workload by supporting the evaluation of essays, especially with regard to language-related criteria. However, due to their tendency to over-rate and their remaining issues to capture the content quality, the models require further refinement.

CCS Concepts

• **Computing methodologies** → *Natural language processing*; • **Applied computing** → **Education**.

Keywords

Large Language Models, Automated Essay Scoring, Learning Analytics, Education

1 Introduction

The correction and evaluation of student texts is a tedious yet crucial process for teachers, especially in subjects where a significant

amount of assessment-relevant text is produced. One of the relatively few studies of teachers' working hours in German schools revealed, for example, that correcting students' texts is enormously time-consuming [34]. The study found that, after teaching itself (29%) and the preparation and follow-up of lessons (22%), proofreading (including feedback to content etc.) accounted for the third-highest proportion of working time for secondary school teachers, at 14%. Notably, the study did not differentiate between subjects, hence teachers of less correction-intensive subjects were also included in this proportion. For example, a typical German teacher with 25 students, three German classes, and four class tests per year — plus one practice essay per test — would need to correct around 600 essays annually, averaging more than two per working day — and that's just for the subject German.

Despite the time and effort teachers invest in correcting student texts, the feedback provided is often delayed, and extremely heterogeneous [32]. Currently, Germany, like many other countries, is also struggling with a systematic shortage of teachers [27], which is expected to worsen in the future. In this context, supporting time-consuming tasks like essay scoring is crucial, and technical solutions can offer valuable assistance to alleviate the workload on teachers.

The rapid advancements of Artificial Intelligence (AI) and Large Language Models (LLMs), such as GPT-4 and o1 by OpenAI [38, 39], LLaMA 3 by Meta [1], or Mistral AI [23] open various new possibilities, also within the educational context [24]. These models possess the capability to process, analyze, and generate natural language with high proficiency, offering the potential to enhance teaching and learning processes. For instance, LLMs can support lesson preparation through automated question generation [8], facilitate teacher collaboration via conversational AI [21], and support the evaluation and feedback generation for student errors [7]. Moreover, the abilities of LLMs extend across various educational domains, including language learning [33], mathematics [37], and life science education [6]. By integrating these advanced AI tools, educators can devote more time to other important tasks, enhance the quality and consistency of student assessments, and foster more interactive and personalized learning experiences.

The challenges mentioned above, coupled with ongoing advancements in AI, raise the question of whether LLMs can serve as efficient alternatives or at least provide valuable support in the evaluation of student essays. Although AI-generated feedback is progressively being integrated into educational applications, there remain substantial gaps in our understanding of the quality and effectiveness of this feedback. While some initial approaches employ AI to provide feedback on student texts [20, 48] and others focus on Automated Essay Scoring (AES) [41, 45], a **detailed, criteria-based evaluation** is still missing. Previous research, mostly evaluating texts based on a holistic score [41, 46] or concentrating on few general and task-agnostic criteria [30, 35], often falls short of fully capturing the complexity and details inherent in student writing, particularly within the context of German language lessons. Also, the limited availability of suitable data resources and the **uncertainty around a clear and consistent ground truth**—typically based on teacher evaluations, which are often subjective and require multiple raters—arise as further challenges. For instance, the widely used ASAP dataset [19] relies on only two raters for most of the essays. While this setup is standard for many datasets, it could limit the robustness of the annotations when evaluating highly subjective dimensions like essay quality. This uncertainty also extends to the ratings of automated essay scoring system. Only few studies have examined the **consistency of LLMs** across multiple runs when evaluating a text [18]. Finally, existing research lacks a **qualitative analysis of the reasoning processes** of LLMs to improve the understanding of which essays aspects have an influence on the model assessment.

To leverage LLMs to their full potential in educational settings, we address these gaps by evaluating the quality of open- and closed-source LLMs in detail by assessing student text based on ten predefined language- and content-related criteria. Therefore, we conducted a user study involving 37 German teachers from different schools and educational levels, collecting on average 5.45 ($SD = 0.92$) ratings per 20 German real-world student essays. Specifically, we compare the human ratings with those generated by GPT-3.5, GPT-4 [38], o1 [39], LLaMA 3-70B [1] and Mixtral 8x7B [23], thoroughly analyzing their strengths and limitations across different evaluation categories and their reasoning process through inter-category correlation examination. We underpin our results by including LLM reliability measurements through comparing repeated assessments. Our analysis adopts a multidimensional approach, extending beyond the mere holistic score to examine biases and performance variations in different assessment aspects and gain in-depth insights into the capabilities of the tested models.

This comprehensive evaluation aims to understand how well LLMs align with human evaluations and to identify areas where they excel or require improvement. To achieve these goals, we target the following research questions:

- **RQ1:** How reliably do open-source and closed-source LLMs perform in essay evaluation, and how do their assessments correlate with real-world teacher evaluations of German-language texts?
- **RQ2:** What are the strengths and limitations of using LLMs to assess multidimensional aspects of essay quality beyond providing a basic holistic score?
- **RQ3:** How do different evaluation criteria influence the reasoning process for essay scores of humans and LLMs?

2 Related Work

2.1 Text assessment

When evaluating student texts, a scale can be established with 'holistic' at one end and 'analytical' at the other [47]. This distinction relates to the nature of the judgment process: holistic approaches typically rely on implicit criteria that are difficult to articulate, while analytical methods are guided by explicit criteria, often structured in detailed rubrics. Despite the scarcity of studies—particularly regarding the assessment of German school texts—research suggests no clear advantage for either approach regarding the statistical quality criteria [17, 47]. Both holistic and analytical assessments are subject to variability between different evaluators [9, 47], though there is also substantial agreement between the judgments derived from each approach [43]. Analytical assessments tend to be more rigorous and less prone to judgment errors [47]. Note, however, that the study situation for German texts is very sparse and that the

various studies were carried out in completely different age groups and text types in different years and are therefore not comparable. From the learners' perspective, criteria-based feedback is key to improving their writing skills, therefore, in this study, we focus on analytical feedback on student texts.

2.2 LLMs for text assessment

Traditional Automated Essay Scoring. Automated Essay Scoring (AES) techniques have been developed since 1966 [45]. Primarily, they exploited statistical features to analyze the text [25]. One notable advancement was the development of Coh-Metrix, which assesses text cohesion using a wide range of linguistic features, providing deeper insights into textual quality [16]. With the introduction of deep learning methods, it became possible to concentrate on more advanced syntactic and semantic features and to analyze the style and content of a text [4], applying LSTMs or Transformer-based models like BERT [12]. For example, BERT has been employed to extract features to train a regression model [13], directly output a class label [50, 55], or combined with an Bi-LSTM to assess essays [5]. Further, hand-crafted features in combination with LSTM or BERT structures have been shown to boost the performance [52]. However, models up to 2022 focus more on language than on content and do not emphasize cohesion and coherence of the essays enough [45]. Also, many works rely on statistical features to determine one holistic final score and concentrate on English data.

Short Answer Grading on German data. To adapt for German essays, [46] applied BERT, focusing on Automatic Short Answer Grading (SAG) (with less than 100 words on average). They achieved a Pearson correlation coefficient of .75, indicating a strong alignment between the model's predictions and human ratings. Similarly, [41] examined the inter-rater reliability between SBERT and human ratings for German corpora, also concentrating on SAG. Although they simplified the task by reducing the grading to a binary decision and training the model on a specific dataset, the transformer model yielded an average precision of only 71.4%. Unlike free-text essay scoring, however, the evaluation of short answers typically relied on a predefined ground truth for each question. This distinction raised concerns about the suitability of transformer models for more complex tasks such as the assessment of student texts using expert ratings. There is an absence of studies focusing on the assessment of standard German essays that go beyond holistic scoring.

Applying GPT models on English Essay data. The rise of LLMs in 2022, has opened up new opportunities for automatic essay grading. One of the first studies was conducted by [11], which compared the ratings of three teachers on 400 English text fragments (averaging 150 words) with the ratings of GPT-3 [40]. The three teachers and the model rated the texts – half written by humans and half generated by GPT-2 [44] – on a five-point Likert scale in four categories: grammaticality, cohesion, sympathy, and relevance. The ratings of the teachers and the model only correlated strongly for the criterion relevance, while the other criteria showed a weak positive correlation. [30] leveraged GPT-3 to evaluate over 12,000 essays from the TOEFL11 dataset [10] across four dimensions: task response, coherence and cohesion, lexical resources, and grammatical range.

They compared GPT-3's performance with 45 linguistic features and discovered that while the linguistic features alone outperformed GPT-3, the combination of both approaches yielded the best results. Specifically, GPT-3 achieved an accuracy of 54% compared to professional human evaluators but demonstrated an 89% agreement within a deviation of 1 to 2 points. [14] reported similar findings with GPT-3.5, showing perfect agreement with human raters on only 30% of exams. However, GPT-3.5 maintained a 70% accuracy within a 10% range around the true score, indicating reasonable but not perfect alignment with human judgments.

In the context of discourse coherence analysis, [35] identified a 56% exact match and a 97% adjacent match between human evaluators and GPT-4 [38]. Additionally, both [2] and [42] employed GPT-4 for open-ended question grading. [2] found a strong correlation between human and LLM scores, while [42] demonstrated that the model's feedback closely resembled expert feedback. [28] compared GPT-3.5 and GPT-4 for automated scoring in science education, revealing that prompting strategies such as few-shot learning and chain-of-thought (CoT) [54] can enhance accuracy.

All these studies rely on closed-source GPT models and focus exclusively on English text data. Moreover, many provide only holistic scores or assess a limited set of artificial criteria. There is a notable lack of detailed analyses exploring how LLM-generated scores differ from human judgments across multidimensional criteria.

Applying open-source LLMs on Essay data. [49] investigated the use of open-source LLMs for generating feedback. They evaluated different prompting strategies for zero- or few-shot learning to determine how well Mistral 7B [22] generates essay feedback. Comparing different prompting strategies appears promising for future research. The study showed that combining AES with feedback generation could improve scoring performance, although the overall impact of AES on feedback quality was minimal. Both, LLM-based scores and manual scores, were used to evaluate the usefulness of the feedback. However, using another LLM to evaluate a feedback-generating LLM raises concerns, such as the risk of perpetuating the biases inherent in the models and the lack of the nuanced understanding of a human expert, such as a teacher. Furthermore, the study did not report the qualifications or backgrounds of the 12 human raters, which is a serious omission. Understanding these raters' expertise is essential to assessing the quality of manual feedback rating. It raises the question of how amateur raters might differ from an LLM in terms of feedback quality.

Acknowledging the limited research in open-source LLMs for automated essay scoring, the scarcity of non-English datasets, and the absence of multidimensional, didactic-based evaluation criteria, our study aims to conduct a comprehensive evaluation of various closed- and open-source LLMs in multidimensional essay assessment and gather new insights into their reasoning processes. In addition, almost no study uses authentic teacher ratings in combination with real learner texts, which impairs the ecological validity of previous studies.

3 Methodology

In our study, we aim to analyze the performance of LLMs in evaluating student texts according to ten pre-defined criteria. In the following, we describe the details of the study, including the essays

and scoring criteria used, the participants, the application of the LLMs, and the metrics employed for analysis.

3.1 Student Essay Dataset

The text corpus comprises $N = 20$ real-world student texts from pupils in Year 7 ($n = 10$) and Year 8 ($n = 10$) at two secondary schools (one "Gymnasium" and one "Realschule") in Germany. These essays were written as part of a performance assessment. Pupils in the seventh grade wrote a narrative with descriptive elements based on the ballad "The Sorcerer's Apprentice" (in German "Der Zauberlehrling") by J. W. Goethe. Those in the eighth grade wrote narratives inspired by two paintings by Edward Hopper ("Nighthawks" and "Gas"). To ensure representativeness despite a small corpus, we selected narratives since they are taught in all German school types.

3.2 Teacher Essay Scoring

The 20 student essays were presented to $N = 37$ teachers with the task to read and assess the texts according to 10 pre-defined criteria using a six-point Likert scale, ranging from *Not true at all* to *Fully applies*. The scale was selected to align with the German grading system, which employs six levels, enabling teachers to provide consistent and familiar ratings. In addition, the even number of possible choices eliminates the risk of a tendency towards the center [53].

Since no universally standardized criteria exist for essay evaluation for any type of text in German, the criteria for this study were derived from real-world teacher feedback specific to these essay types and were later refined and formalized by an expert in German didactic. The advantage of this approach, apart from the increase in ecological validity, is that teacher are familiar with the categories, which in turn has a positive influence on the reliability of their ratings. Half of the resulting categories asses the content (i.e., *Introduction*, *Main Part*, and *Conclusion*), while the others focus on the writing style and language (e.g., *Verbal images* and *Descriptive elements*, *Literal speech & inner monologue* and *Spelling & punctuation*). Further, we added one category (*Overall*) to ask for an holistic rating, as detailed in Table 1. The teachers submitted their assessments anonymously via an online portal. Each teacher was asked to rate three texts; however, there are missing values, resulting in a total of 1,090 human ratings across the 20 texts and the 10 assessment categories. Each essay received between three and seven ratings from different teachers, with an average of 5.45 ± 0.92 ratings per essay. We employ the average of the teacher ratings as our ground truth per essay and evaluation criteria.

3.3 LLM Essay Scoring

To automatically evaluate student essays based on the ten pre-defined criteria, we selected different LLMs, to compare the performance of various foundation models. For closed-source models, we used GPT-3.5 (`gpt-3.5-turbo-0125`), GPT-4 (`gpt-4o-2024-05-13`) [38] and o1 (`o1-preview`) [39] as representative examples. These GPT models were integrated into our evaluation pipeline via the OpenAI API. For open-source LLMs, we chose LLaMA 3-70B [1] and Mixtral 8x7B [23]. Initial experiments included smaller variants of these models; however, their performance was inferior to their

larger counterparts. Therefore, the smaller models were excluded from further analysis. Both open-source models were executed locally in half-precision using two NVIDIA A100 80GB GPUs to ensure efficient processing.

All LLMs were prompted using a zero-shot approach, instructing them to evaluate one text at a time. For each essay, a new conversation was initiated to maintain the independence of evaluations. The prompt structure was designed to align with the predefined evaluation criteria, ensuring consistency across all model assessments. This methodology allowed us to systematically compare the performance of different LLMs in a controlled and unbiased manner.

```
     You are a teacher. Analyze the essay written by a
13-year-old child according to the given criteria. Return a
  scalar number from 1 to 6 for each criteria. 1 means the
  criteria is not fulfilled at all, 6 means it is completely
accomplished. Return only a JSON. ## Criteria = {criteria};
            ## Essay = {text}; ## Rating =
```

Figure 2: Standardized zero-shot prompt employed across all LLMs to ensure a fair comparison of their essay evaluation performance.

The prompt, illustrated in Figure 2, specifies the model's role as a teacher, provides the context of an essay written by a 13-year-old student, outlines the specific task of analyzing the essay according to predefined criteria, and defines the desired output format as JSON. This prompt was consistently used throughout all experiments to ensure uniformity. While variations in input prompts can lead to different results [49], specific prompt engineering was not the focus here.

To analyze reliability and capture the stochastic nature of models, each text was evaluated ten times by each LLM at a temperature of 0.7. The mean of these ten assessments represents the average rating assigned by the LLMs, analogous to the average of multiple human ratings for the same text. For open-source models, we filtered out any missing values or predictions that fell outside the predefined range ($N = 169$). In contrast, the GPT models consistently returned outputs in a valid format, ensuring the integrity of the collected data. A total of $N = 9,931$ LLM ratings were collected.

3.4 Analysis

To investigate our research questions, we systematically analyzed the assessments generated by both human raters and LLMs. **RQ1** addresses the reliability and quality of the ratings. To answer this, we compared the different runs of each model on the same text by calculating inter-class correlation, treating each run as an individual rater (see Section 4.1). For the following sections, we always consider the average of the ten runs as the final rating. In Section 4.2, we compared the assessments of the LLMs to those of real-world teachers using the Spearman correlation coefficient to identify similarities. For **RQ2**, we examined the overall holistic scores alongside all multidimensional aspects, including both language- and content-related criteria (see Section 4.3). By comparing the distribution of

No.	Title	Type	Description
1	Headline	C	The headline creates suspense and is appropriately chosen.
2	Introduction	C	The introduction is interesting and builds up the suspense.
3	Main part	C	The main part builds up the suspense convincingly and skillfully and contains a turning point.
4	Conclusion	C	The conclusion rounds off the story and brings all the narrative strands to a close.
5	Verbal images & descriptive elements	L	The choice of words is vivid and linguistic images / descriptive elements are used.
6	Literal speech & inner monologue	L	Literal speech and inner monologue are used purposefully.
7	Plot logic	C	The plot is structured without logical breaks.
8	Expression & sentence structure	L	The expression and sentence structure are varied and convincing.
9	Spelling & punctuation	L	The spelling and punctuation are secure.
0	Overall judgment	C/L	Overall, the work is...

Table 1: The 10 evaluation criteria, categorized into content-related (C) and language-related (L) aspects, each assessed using a six-point Likert scale.

ratings between LLMs and human raters and applying the Mann-Whitney U test, we identified any differences in how specific criteria are rated by each group. Lastly, **RQ3** explores the factors influencing the reasoning processes behind the evaluations. In Section 4.5, we analyzed inter-category correlations for both human raters and LLMs. This analysis helps us understand which categories have the strongest associations with the overall score, providing insights into the reasoning processes of both humans and LLMs and identifying which criteria most significantly influence the holistic assessment.

4 Results

In the following section, we explore the discrepancies between real-world teacher assessments and open- as well as closed-source LLM-generated evaluations of student texts, aiming to address our three key research questions.

4.1 Reliability of model predictions

We compare multiple runs of the same prompt to assess the reliability of each model to answer RQ1. Consequently, we obtain ten ratings for each data point. Table 2 presents the ICC statistics for all employed foundation models.

	GPT-3.5	GPT-4	GPT-o1	LLaMA	Mixtral
ICC	0.84	0.73	0.80	-0.04	0.01

Table 2: ICC values comparing the inter-rater reliability of each LLM across multiple evaluation runs.

It is evident that the closed-source models demonstrate a fair level of consistency in their ratings across individual runs, as indicated by ICC showing moderate to good reliability (.73 - .84) [26]. In contrast, the open-source models LLaMA 3 and Mixtral, display significant variability, with low values and poor agreement between multiple runs. This inconsistency needs to be taken into account in the subsequent analysis.

4.2 Correlation Analysis Between LLM Evaluations and Teacher Assessments

Continuing our examination of RQ1, we compare the ratings of LLMs and humans by computing the Spearman correlation coefficients r for all evaluation criteria in Table 3. o1 demonstrates the strongest correlation with human ratings, achieving significance in nine out of ten categories. Specifically, it is the only model demonstrating a high, significant correlation of .74 to teacher ratings in the overall category. GPT-4 maintains significant correlation in seven out of ten categories, with a moderate agreement with human raters in most of the categories, while GPT-3.5 reaches this in four categories. This again shows the improvement across subsequent model versions. In contrast, Mixtral shows only weak or non-significant correlations across all criteria. LLaMA 3 exhibits nearly zero correlation to teacher ratings and, in some cases, even negative correlations, underscoring its inconsistency with human evaluations.

Figure 3 visually shows the strong correlation between o1 and human ratings, especially for language-related features such as spelling ($r=0.814$), use of literal speech ($r=.805$), and verbal imagery ($r=.738$). Additionally, the overall rating exhibits a robust correlation despite having an intercept of 2.3. However, two content-related categories — *Introduction*, and *Main Part* — demonstrate only weak or moderate correlations. In these areas, o1 is less effective at distinguishing between higher and lower-quality essays.

4.3 Rating Comparison between LLM and Human Evaluation

To address RQ2, we first compare the distribution of the ratings in the *Overall* category, as shown in Table 4. The human raters exhibit a lower average rating than the GPT models, indicating a stricter assessment approach. The LLaMA 3 model presents a similar average to the human raters. However, the variance in ratings across essays is significantly higher for human raters than for all LLMs, with the LLaMA 3 and Mixtral displaying particularly low variance. This suggests that the open-source models assess essays more uniformly, clustering ratings around the midpoint, whereas human raters and the closed-source models provide a

Category	GPT-3.5		GPT-4		GPT-o1		LLaMA 3-70B		Mixtral 8x7B	
	r	p	r	p	r	p	r	p	r	p
Overall	0.418	0.067	**0.575**	**0.008**	**0.742**	**0.000**	0.091	0.703	0.311	0.182
Headline	-0.005	0.984	0.159	0.504	**0.699**	**0.001**	0.131	0.581	0.252	0.284
Introduction	0.313	0.179	0.325	0.162	0.127	0.594	0.014	0.953	0.348	0.133
Main part	0.279	0.234	0.386	0.092	**0.466**	**0.038**	-0.094	0.694	0.376	0.103
Conclusion	**0.450**	**0.046**	**0.453**	**0.045**	**0.714**	**0.000**	-0.288	0.218	0.011	0.964
Verbal image	**0.483**	**0.031**	**0.680**	**0.001**	**0.738**	**0.000**	-0.064	0.789	0.146	0.538
Literal Speech	0.138	0.561	**0.521**	**0.019**	**0.805**	**0.000**	-0.262	0.264	0.375	0.103
Plot logic	0.425	0.062	**0.585**	**0.007**	**0.608**	**0.004**	-0.032	0.893	0.442	0.051
Expression	**0.520**	**0.019**	**0.626**	**0.003**	**0.675**	**0.001**	0.177	0.455	0.211	0.372
Spelling	**0.728**	**0.000**	**0.846**	**0.000**	**0.814**	**0.000**	0.406	0.076	0.005	0.984

Table 3: Spearman correlation coefficients (r) and the corresponding p-values comparing human ratings with those of the five LLM-models. Significant correlations are highlighted in bold.
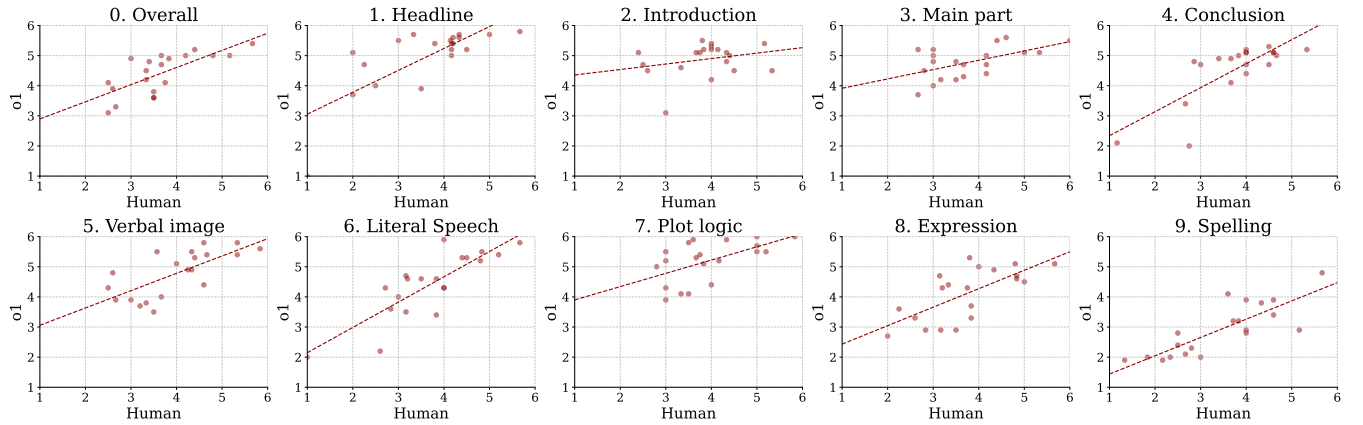

Figure 3: Correlation between o1 and human ratings across all evaluation criteria. The red line represents the linear least-squares regression between the two sets of ratings.

wider differentiation between essay qualities. Additionally, GPT-4, o1, and Mixtral provide significantly higher ratings than humans. The low rater correlation of both LLaMA 3 and Mixtral (Section 4.1) contributes to these observations by resulting in more average values and reduced variance. This highlights a limitation in the consistency and alignment of certain LLMs with human evaluative standards.

4.4 Analysis of Rating Distributions Across Evaluation Categories

To gain further insights into the differing ratings, we compare the average ratings for each criterion in Figure 4. Generally, the GPT models exhibit higher mean ratings across all individual criteria, whereas human raters tend to be stricter in their evaluations. LLaMA 3 displays average ratings similar to those of the human raters. In contrast, the Mixtral model consistently shows significantly higher average ratings across all categories. Consequently, the differences observed in the overall ratings between LLMs and

humans are reflected in each category's judgments. This consistency indicates that the final ratings align with the more detailed evaluations of human raters and LLMs.

It is important to note the variances presented in Table 5. While the LLaMA 3 models exhibit average ratings comparable to human raters, their variances are considerably smaller, consistently clustering around the midpoint. Similarly, the Mixtral model also demonstrates reduced variance, aligning with the pattern observed in LLaMA 3. In contrast, GPT-3.5, o1, and human raters display higher variances, indicating a broader differentiation in their evaluations. The low ICC values for both LLaMA 3 and Mixtral models contribute to these observations, as their limited agreement among ratings results in more average values and diminished variability.

To explore the mean value comparisons from Table 4 in more detail, Table 5 also presents the Mann-Whitney U test p-values, which compare the distribution of ratings across each criterion between humans and the corresponding LLM. There are no significant differences between human and GPT-4 ratings for *Conclusion*, and for all the language-related criteria. Similarly, o1 only exhibits a non-significant difference for *Spelling & Punctuation* and *Expression*, both

	Human	GPT-3.5	p	GPT-4	p	GPT-o1	p	LLaMA 3	p	Mixtral	p
Overall	3.65 (0.86)	3.85 (0.81)	0.30	**4.34 (0.66)**	**0.02**	4.41 (0.68)	**0.01**	3.55 (0.41)	0.97	**4.60 (0.41)**	**0.00**

Table 4: Overall average evaluations of all essays (standard deviations), including p-values from the Mann-Whitney U test comparing the distribution of the LLM scores to human scores. Significant differences are indicated in bold.
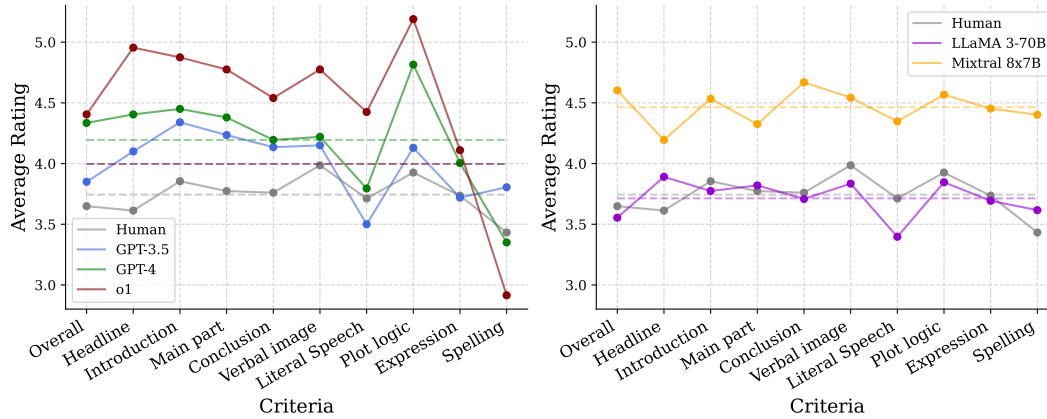

Figure 4: Average ratings for each evaluation criterion, illustrating the differences in assessments between human raters and closed-sourced LLMs (left side) and open-sources LLMs (right side).

Category	Human	GPT-3.5	p	GPT-4	p	GPT-o1	p	LLaMA 3	p	Mixtral	p
Headline	3.61 (1.16)	4.10 (0.82)	0.10	**4.40 (0.60)**	**0.00**	**4.96 (1.12)**	**0.00**	3.89 (0.45)	0.75	4.19 (0.59)	0.09
Introduction	3.85 (0.79)	4.34 (0.91)	0.07	**4.45 (0.47)**	**0.01**	**4.87 (0.55)**	**0.00**	3.77 (0.43)	0.39	**4.53 (0.53)**	**0.00**
Main part	3.77 (0.94)	4.24 (1.14)	0.17	**4.38 (0.57)**	**0.01**	**4.78 (0.53)**	**0.00**	3.82 (0.51)	0.53	**4.33 (0.51)**	**0.02**
Conclusion	3.76 (0.93)	4.14 (0.91)	0.13	4.20 (0.67)	0.11	**4.54 (0.96)**	**0.00**	3.71 (0.55)	0.45	**4.67 (0.49)**	**0.00**
Verbal image	3.99 (0.95)	4.15 (1.08)	0.47	4.22 (0.74)	0.32	**4.78 (0.77)**	**0.01**	3.83 (0.59)	0.63	**4.54 (0.51)**	**0.05**
Literal Speech	3.71 (1.07)	3.50 (1.09)	0.56	3.80 (0.48)	0.79	**4.42 (1.08)**	**0.03**	3.40 (0.38)	0.27	**4.35 (0.57)**	**0.02**
Plot logic	3.93 (0.88)	4.13 (1.17)	0.66	**4.81 (0.59)**	**0.00**	**5.19 (0.68)**	**0.00**	3.85 (0.51)	0.83	**4.57 (0.47)**	**0.01**
Expression	3.74 (0.97)	3.72 (0.98)	0.92	4.01 (0.68)	0.34	4.11 (0.86)	0.24	3.69 (0.45)	0.96	**4.45 (0.50)**	**0.01**
Spelling	3.43 (1.16)	3.80 (1.05)	0.36	3.35 (0.91)	0.83	2.91 (0.86)	0.14	3.62 (0.72)	0.66	**4.40 (0.47)**	**0.00**

Table 5: Average criteria-based evaluations (SD), including p-values from the Mann-Whitney U test comparing the distribution of human and LLM ratings for each category. Significant differences between human and LLM average scores are highlighted in bold.

language-categories. Since LLMs are trained on vast amounts of text data encompassing various writing styles and formalities, they are optimized for recognizing patterns and stylistic elements. Consequently, these surface-level aspects of a text, including grammar and sentence structure, can be efficiently analyzed and evaluated by LLMs in a manner similar to humans. Remarkably, they accurately assess learner texts despite likely limited exposure to student writings during training, demonstrating their effective generalization for educational writing evaluation.

In contrast, the discrepancy between human ratings and GPT-4 and o1 is significant for criteria *Heading*, *Introduction*, *Main Part*, and *Plot Logic*, all of which are content-related categories. This gap may arise from several factors. Although LLMs can generate coherent

texts, at the moment they still lack the deep semantic understanding that humans possess. This limitation manifests in flawed logical reasoning concerning contextual details and maintaining logical consistency. Additionally, during the training process, model biases can be inherited and learned by the LLMs, resulting in milder ratings. Notably, their are differences between model versions, as previously indicated in Table 4 and Figure 4, which is surprising given that newer model versions typically outperforms their predecessor in complex tasks, which is expected to bring it closer to human evaluative standards. These findings highlight the ongoing challenges in aligning LLM assessments with human judgment, particularly in content-heavy criteria.

4.5 Impact of Evaluation Categories on Overall Essay Scoring

To compare how strongly individual evaluation categories impact the overall rating for teachers and LLMs (RQ3), we correlate single criteria scores with the overall rating for teachers and each model. Overall, it is noteworthy that the correlations between all features of GPT-3.5 and GPT-4 are higher than those for human judgments. This can be attributed to the auto-regressive nature of LLMs, where each new token generation is conditioned on the previous output, implicitly incorporating prior evaluations into the context. This effect is particularly pronounced in the smaller GPT-3.5 model, where correlations between ratings are nearly always $\rho \geq 0.8$. The high inter-criteria correlations may indicate that the assessment of each criterion relies more on preceding ratings than solely on the original text.

In contrast, o1 exhibits fairly lower correlations between criteria, which may stem from its larger number of parameters and enhanced context understanding, allowing it to focus more on the input text than previous ratings. Compared to the previous GPT models, human raters and o1 demonstrate lower inter-criteria correlations, suggesting that their evaluations are less influenced by preceding ratings and more independently based on each criterion's merits.

Conversely, LLaMA 3 and Mixtral consistently show low correlations across criteria, indicating a distinct evaluation pattern and potentially a less integrated evaluation strategy compared to closed-source models.

When examining which evaluations criteria seem to highly impact overall scores, both human raters and LLMs agree that criteria *Plot Logic* and *Expression & Sentence Structure* are highly relevant to the overall grade, whereas criterion *Heading* is less relevant. Major differences are observed in how criterion *Main Part* influences the overall grade for teachers compared to GPT-4, while o1 is able to include this information more strongly. Specifically, for teachers, the *Main Part* is a major determinant of the overall assessment, whereas GPT-4 places greater emphasis on criterion *Spelling & Punctuation* and *Verbal image*, which is not as decisive for human evaluators. This discrepancy highlights a potential risk associated with using LLMs for assessing student texts: the models may inherently focus more on linguistic surface features, such as *Spelling & punctuation*. Consequently, these surface-elements could be weighted more heavily in the overall assessment than is appropriate, potentially undermining the evaluation of deeper content quality. In addition, *Main Part* and *Plot Logic* are particularly influential criteria for the teachers, but precisely here Table 3 shows weak correlations and Table 5 significant differences between GPT-4, o1 and the humans. This shows one of the major challenges we are currently facing in improving machine-generated feedback.

5 Discussion

This study aimed to evaluate the performance and reliability of open-source and closed-source LLMs in assessing student essays, specifically focusing on German-language texts.

5.1 Closed-source LLMs provide reliable essay assessments

Our analysis revealed that closed-source models, particularly o1, exhibit higher reliability when run multiple times and stronger correlations with human assessments compared to open-source models like LLaMA 3 and Mixtral. The high reliability of GPT models aligns with previous findings in the literature [18]. o1 demonstrated strong correlations in eight out of ten evaluation categories, aligning closely with human judgments in language-related aspects.

Conversely, open-source models like LLaMA 3 and Mixtral showed minimal to no correlation with human ratings, often rating essays around the midpoint with low variance. This limited differentiation suggests that these models are less effective in distinguishing between varying levels of essay quality, likely due to their lower ICC scores indicating poor internal consistency. This inconsistency in ratings, when run multiple times, makes them unsuitable for employment in real-world scenarios. [49] find a similar lack of alignment with human evaluations with LLaMA 2 [51], while they were able to increase the performance of Mistral by applying prompt engineering techniques like CoT prompting [54].

5.2 LLMs align best with Human Ratings in Language-Related Criteria

The comparison of overall ratings on individual criteria revealed that GPT-4, o1, and Mixtral generally provide higher average scores than human raters, which is in line with previous research [2, 31]. The lower variance in overall scores observed in LLaMA 3 and Mixtral models indicates a tendency to rate essays uniformly, likely due to averaging highly inconsistent values of several assessments, which limits their ability to differentiate effectively between high and low-quality writing. This finding is consistent with our initial observation of low internal consistency for these models, suggesting that they lack detailed evaluation capabilities.

Moreover, the Mann-Whitney U tests indicated that the distribution of GPT-4 ratings aligns closely with human judgments in language-related criteria but diverges significantly in content-related aspects, which have the highest influence on the overall grade for real-world teachers. This indicates a better alignment in broadly applicable linguistic features but a bigger discrepancy for content criteria, which are highly dependent on the context (i.e., grade level and text type) and on the particular standards and conventions of the German school system. This may stem from the model's lack of exposure to similar data during pre-training. Further alignment through fine-tuning or few-shot learning might mitigate this issue by providing the models with a frame of reference. Otherwise, deploying these tools in real-world educational settings could lead to inconsistent assessments and reduced usability.

5.3 Inter-Criteria Correlations emphasize Consistent Reasoning in GPT Models

The correlation analysis revealed that GPT models exhibit higher inter-criteria correlation compared to human raters. This phenomenon can be attributed to the auto-regressive nature of LLMs, where each generated token is influenced by preceding outputs, thereby creating a more integrated and consistent reasoning process across

	Human Overall	GPT-3.5 Overall	GPT-4 Overall	o1 Overall	LLaMA 3-70B Overall	Mixtral 8x7B Overall
Overall	1.00	1.00	1.00	1.00	1.00	1.00
Headline	0.71	0.86	0.75	0.68	0.37	0.14
Introduction	0.73	0.94	0.93	0.61	0.41	0.52
Main part	0.88	0.95	0.83	0.88	0.71	0.53
Conclusion	0.74	0.93	0.75	0.66	0.55	0.38
Verbal image	0.78	0.94	0.93	0.85	0.63	0.50
Literal Speech	0.81	0.95	0.76	0.38	0.71	0.49
Plot logic	0.90	0.93	0.93	0.81	0.80	0.62
Expression	0.87	0.95	0.96	0.95	0.71	0.65
Spelling	0.61	0.94	0.89	0.89	0.67	0.78

Figure 5: Inter-criteria correlation matrix illustrating the strength of relationships between individual evaluation categories for each model. Light blue indicates weak correlations, while dark blue marks strong correlations.

different evaluation dimensions. Specifically, GPT-3.5 and GPT-4 showed strong correlations ($\rho \geq 0.86$ resp. $\rho \geq 0.75$) between criteria, suggesting that its assessments are highly dependent on prior ratings. While this may enhance consistency, it also raises concerns about the model's ability to independently evaluate each criterion based solely on the original text.

In contrast, o1 demonstrated lower and more diverse correlations than GPT-3.5 and GPT-4, likely due to its larger parameter set and enhanced context understanding, which allow for a more balanced focus between the input text and prior evaluations. This results in a reasoning process that better reflects human evaluative practices, particularly in content-related categories. The open-source models LLaMA 3 and Mixtral, however, maintained weaker correlations across criteria, indicating a fragmented evaluation approach that lacks the cohesive reasoning seen in GPT models. These patterns suggest that the architectural and training differences between closed-source and open-source models significantly impact their evaluation strategies and reliability.

While for humans content-related criteria like main part and plot logic have the highest influence on the overall rating, GPT-4 sets a greater focus on language aspects like expression and spelling. This discrepancy points, again, to a basic limitation where LLMs may prioritize surface-level linguistic features over deeper content analysis, potentially skewing the overall assessment towards aspects like spelling and punctuation rather than substantive content quality.

5.4 Implications for Automated Essay Assessment

The findings of this study underscore both the potential and the limitations of using LLMs for AES. GPT-4 and o1, despite their high alignment with human ratings in language-related criteria, exhibit a bias towards higher scores. This highlights the need for careful calibration and potential alignment strategies, such as top-down prompting or bottom-up fine-tuning on domain-specific datasets, to mitigate biases and enhance the ability to assess essays more accurately.

Moreover, the poor performance of open-source models like LLaMA 3 and Mixtral suggests that these models are not yet ready for reliable use in educational settings without significant improvements in their consistency and alignment with human standards. The low ICC scores and minimal correlation with human ratings indicate that these models lack the necessary reliability for detailed, multidimensional assessments required in educational contexts. When considering real-world applications, it is crucial that distinct models or variations in prompts can lead to significant differences in outcomes. This calls for a thorough evaluation of these techniques to ensure fairness and uphold ethical standards in their deployment.

5.5 Trust and Perceived Usefulness

[56] found in their meta-study on the perceived usefulness of automated writing assessments with Chinese students that students' trust in the automated system is an important factor in the perceived usefulness of the system. Moreover, [36] showed that it is important for teachers to understand how the decisions are made to increase trust in a system. Additionally, [29] found that emotional trust in products could be negatively influenced by the 'AI' label. Against this background, it is important not to waste the potential that LLMs have for the assessment of student texts by overestimating their functionality while their development is still at the beginning.

Building on these previous findings, the high correlation and reliability of GPT-4 and o1 could foster greater trust among educators and students, enhancing the perceived usefulness of such systems. However, transparency in model limitations and ongoing efforts to align LLM assessments with human evaluative standards are essential to implementing the full potential of these technologies without limiting user confidence.

5.6 Limitations and Future Directions

This study has several limitations that should be acknowledged. First, while no prompt engineering was employed in this research, future work on open-source models could benefit from sophisticated

strategies such as Chain-of-Thought Engineering [54], as demonstrated in [49]. The novel o1 series automatically incorporates CoT reasoning, which likely contributes to its superior performance. Implementing these techniques may enhance the performance and alignment of LLMs with human ratings by enabling more detailed and context-aware evaluations.

The scope of this study was limited to a specific set of models, namely GPT-3.5, GPT-4, o1, LLaMA 3, and Mixtral. Incorporating other models such as Claude [3] or Gemini [15] in future research would offer a more comprehensive evaluation of LLM performance across different architectures and training paradigms. Furthermore, this study focused solely on one type of essay. Exploring other essay formats, such as argumentative essays, and using a more extensive and diverse dataset would enhance the generalizability and robustness of the findings. Different essay types may present unique challenges and require distinct evaluative criteria, providing a more holistic understanding of LLM capabilities in automated assessment. Also, the inherent variance observed in multiple runs of LLMs, specifically for open-source models, suggests that real-world applications should incorporate mechanisms to aggregate multiple scores to mitigate the influence of outliers and enhance the robustness. Additionally, the absence of a gold standard due to the inherent variability among human raters poses a challenge for automated systems. Future studies should investigate methods to account for the fuzzy nature of human evaluations and integrate them into LLM training and assessment frameworks.

Finally, we observed a clear trend in OpenAI's closed-source models: each new version shows improved reliability and a stronger correlation with human assessments, which was also shown by [28]. Given the rapid advancements in this field and our findings, we anticipate continued enhancements in LLMs, making them increasingly effective tools for supporting automated essay evaluation in educational settings.

6 Conclusion

In this study, we evaluated open-source and closed-source Large Language Models in assessing German student essays against human ratings, highlighting the strengths and limitations of multidimensional evaluation by LLMs. The novel o1 model demonstrated high reliability and strong correlations with human ratings in language-related criteria, though it tended to assign higher overall scores. In contrast, open-source models like LLaMA 3 and Mixtral showed low variance and weak correlations, limiting their effectiveness for educational assessments. These findings suggest that while o1 holds promise as an assistive tool for teachers, further refinement is necessary to enhance content evaluation and ensure alignment with human standards, especially when the goal is not just to award points on pre-determined criteria, but to provide feedback that is conducive to learning, showing students specifically what and how to improve their text.

References

[1] AI@Meta. 2024. Llama 3 Model Card.

[2] Hani Alers, Aleksandra Malinowska, Gregory Meghoe, and Enso Apfel. 2024. Using ChatGPT-4 to Grade Open Question Exams. In *Advances in Information and Communication*, Kohei Arai (Ed.). Springer, Springer Nature Switzerland, Cham, 1–9.

[3] Anthropic. 2023. Claude. https://www.anthropic.com/

[4] Xiaoyu Bai and Manfred Stede. 2023. A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education* 33, 4 (2023), 992–1030.

[5] Majdi Beseiso, Omar A Alzubi, and Hasan Rashaideh. 2021. A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education* 33 (2021), 727–746.

[6] Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. 2024. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. arXiv:2401.00832

[7] Arne Bewersdorff, Kathrin Seßler, Armin Baur, Enkelejda Kasneci, and Claudia Nerdel. 2023. Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence* 5 (2023), 100177.

[8] Shravya Bhat, Huy Anh Nguyen, Steven Moore, John C Stamper, Majd Sakr, and Eric Nyberg. 2022. Towards Automated Generation and Evaluation of Questions in Educational Domains.. In *EDM*, Antonija Mitrovic and Nigel Bosch (Eds.). International Educational Data Mining Society, Durham, United Kingdom, 701–704.

[9] Peter Birkel and Claudia Birkel. 2002. Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, 49, 3 (2002), 219–224.

[10] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. *TOEFL11: A Corpus of Non-Native English*. Technical Report. Educational Testing Service.

[11] Cheng-Han Chiang and Hung-yi Lee. 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[13] Afrizal Doewes and Mykola Pechenizkiy. 2021. On the Limitations of Human-Computer Agreement in Automated Essay Scoring.. In *The 14th International Conference on Educational Data Mining (EDM21)*. International Educational Data Mining Society, Paris, France, 475–480.

[14] Jonas Flodén. 2024. Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal* 00 (2024), 1–24.

[15] Google Gemini Team. 2024. Gemini: A Family of Highly Capable Multimodal Models.

[16] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 2 (2004), 193–202.

[17] Jürgen Grzesik and Michael Fischer. 1984. *Was leisten Kriterien für die Aufsatzbeurteilung?: Theoretische, empirische und praktische Aspekte des Gebrauchs von Kriterien und der Mehrfachbeurteilung nach globalem Ersteindruck*. Westdeutscher-Verlag, Wiesbaden, Germany.

[18] Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education* 8, 1272229.

[19] Ben Hamner, Jaison Morgan, Iynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring. https://www.kaggle.com/c/asap-aes

[20] Hendrik Haverkamp, Malte Hecht, and Kirsten Schindler. 2024. Lernförderliches Feedback KI-basiert vermitteln. *Der Deutschunterricht* 5 (2024).

[21] Hyangeun Ji, Insook Han, and Yujung Ko. 2023. A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education* 55, 1 (2023), 48–63.

[22] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. *Mistral 7B (2023)*. Technical Report. Mistral AI.

[23] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. *Mixtral of experts*. Technical Report. Mistral AI.

[24] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[25] Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art.. In *Proceedings of the Twenty-Eighth International Joint Conference*

on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, 6300–6308.

[26] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine 15, 2 (2016), 155–163.

[27] Kultusministerkonferenz. 2023. Lehrkräfteeinstellungsbedarf und -angebot in der Bundesrepublik Deutschland 2023 – 2035: Zusammengefasste Modellrechnungen der Länder. Dokumentation 238. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Berlin. Beschluss der Kultusministerkonferenz vom 08.12.2023.

[28] Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. Computers and Education: Artificial Intelligence 6 (2024), 100213.

[29] Dogan Gursoy Mesut Cicek and Lu Lu. 2024. Adverse impacts of revealing the presence of "Artificial Intelligence (AI)" technology in product and service descriptions on purchase intentions: the mediating role of emotional trust and the moderating role of perceived risk. Journal of Hospitality Marketing & Management 0, 0 (2024), 1–23.

[30] Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics 2, 2 (2023), 100050.

[31] Leo Morjaria, Levi Burns, Keyna Bracken, Anthony J Levinson, Quang N Ngo, Mark Lee, and Matthew Sibbald. 2024. Examining the Efficacy of ChatGPT in Marking Short-Answer Assessments in an Undergraduate Medical Program. International Medical Education 3, 1 (2024), 32–43.

[32] Nora Müller, Till Utesch, and Vera Busse. 2023. Qualität statt Quantität? Zum Zusammenhang von Schreibförderungs-und Feedbackpraktiken mit Textqualität unter Berücksichtigung von migrationsbedingter Mehrsprachigkeit. Unt.wiss. Zeits. f. Lernforschung 51, 2 (2023), 169–198.

[33] Sonia Alejandrina Sotelo Muñoz, Giovanna Gutiérrez Gayoso, Alberto Caceres Huambo, Rogelio Domingo Cahuana Tapia, Jorge Layme Incaluque, Oscar Eduardo Pongo Aguila, Juan Cielo Ramírez Cajamarca, Jesus Enrique Reyes Acevedo, Herbert Victor Huaranga Rivera, and José Luis Arias-Gonzáles. 2023. Examining the impacts of ChatGPT on student motivation and engagement. Social Space 23, 1 (2023), 1–27.

[34] Frank Mußmann, Martin Riethmüller, Thomas Hardwig, Stefan Peters, Marcel Parciak, Ilka Charlotte Ohms, and Stefan Klötzer. 2016. Niedersächsische Arbeitszeitstudie 2015 / 2016: Lehrkräfte an öffentlichen Schulen Ergebnisbericht. Technical Report. Georg-August-Universität Göttingen, Kooperationsstelle Hochschulen und Gewerkschaften, Humboldtallee 15, D-37073 Göttingen. https://doi.org/10. 3249/webdoc-3971 Redaktionell überarbeitet im Dezember 2016.

[35] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.). Association for Computational Linguistics, Toronto, Canada, 394–403.

[36] Tanya Nazaretsky, Moriah Ariely, Mutlu Cukurova, and Giora Alexandron. 2022. Teachers' trust in AI-powered educational technology and a professional development program to improve it. British journal of educational technology 53, 4 (2022), 914–931.

[37] Huy A Nguyen, Hayden Stec, Xinying Hou, Sarah Di, and Bruce M McLaren. 2023. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In European Conference on Technology Enhanced Learning, Olga Viberg, Ioana Jivet, Pedro J. Muñoz-Merino, Maria Perifanou, and Tina Papathoma (Eds.). Springer Nature Switzerland, Cham, 278–293.

[38] OpenAI. 2023. GPT-4 technical report. Technical Report. OpenAI.

[39] OpenAI. 2024. OpenAI o1 System Card. Technical Report. OpenAI.

[40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, et al. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems 35 (2022), 27730–27744.

[41] Ulrike Padó, Yunus Eryilmaz, and Larissa Kirschner. 2023. Short-Answer Grading for German: Addressing the Challenges. International Journal of Artificial Intelligence in Education (2023), 1–32.

[42] Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, and Kiev Gama. 2023. Large language models for education: Grading open-ended questions using chatgpt. In Proceedings of the XXXVII Brazilian Symposium on Software Engineering. Association for Computing Machinery, New York, NY, USA, 293–302.

[43] Sanna Pohlmann-Rother, Edgar Schoreit, and Anja Kürzinger. 2016. Schreibkompetenzen von Erstklässlern quantitativ-empirisch erfassen-Herausforderungen und Zugewinn eines analytisch-kriterialen Vorgehens gegenüber einer holistischen Bewertung. Journal for educational research online 8, 2 (2016), 107–135.

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 8 (2019), 9.

[45] Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review 55, 3 (2022), 2495–2527.

[46] Jörg Sawatzki, Tim Schlippe, and Marian Benner-Wickner. 2021. Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. In International conference on artificial intelligence in education technology, Eric C. K. Cheng, Rekha B. Koul, Tianchong Wang, and Xinguo Yu (Eds.). Springer Nature Singapore, Singapore, 65–75.

[47] Pauline Schröter, Hannelore Söldner, Lars Hoffmann, Anja Riemenschneider, Jörg Jost, and Dorothee Wieser. 2022. Wie vergleichbar sind die Bewertungen von Abiturarbeiten im Fach Deutsch? Empirische Studien zu verschiedenen Bewertungsmodellen. In Das unvergleichliche Abitur: Entwicklungen–Herausforderungen–Empirische Analysen, Lars Hoffmann, Pauline Schröter, Alexander Groß, Svenja Schmid-Kühn, and Petra Stanat (Eds.). wbv Bielefeld, Bielefeld, Germany, 213–250.

[48] Kathrin Seßler, Tao Xiang, Lukas Bogenrieder, and Enkelejda Kasneci. 2023. Peer: Empowering writing with large language models. In Responsive and Sustainable Educational Futures, Olga Viberg, Ioana Jivet, Pedro J. Muñoz-Merino, Maria Perifanou, and Tina Papathoma (Eds.). Springer Nature Switzerland, Cham, 755–761.

[49] Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. arXiv:2404.15845 [cs.CL]

[50] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using transformer-based pre-training. In Artificial Intelligence in Education, Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin (Eds.). Springer International Publishing, Cham, 469–481.

[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. Technical Report. GenAI, Meta.

[52] Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In Proceedings of the 28th International Conference on Computational Linguistics, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 6077–6088.

[53] Hester van Herk, Ype H. Poortinga, and Theo M. M. Verhallen. 2004. Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. Journal of Cross-Cultural Psychology 35, 3 (2004), 346–360.

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35 (2022), 24824–24837.

[55] Jin Xue, Xiaoyi Tang, and Liyan Zheng. 2021. A hierarchical BERT-based transfer learning approach for multi-dimensional essay scoring. Ieee Access 9 (2021), 125403–125415.

[56] Na Zhai and Xiaomei Ma. 2022. Automated writing evaluation (AWE) feedback: a systematic investigation of college students' acceptance. Computer Assisted Language Learning 35, 9 (2022), 2817–2842.