Didaktik Deutsch

Afra Sturm, Valentin Unger & Fabian Grünig

Human versus machine: Comparing human and large-language-model assessments of students' writing through benchmark ratings

Abstract

For teachers to effectively use large-language-model(LLM)-based ratings in the formative or summative assessment of texts, it is essential to ensure that such ratings can assess student writing in a valid and reliable manner. This study investigates whether a validated human text-rating procedure (benchmark rating) can be replicated by an LLM-based rating procedure. We tested the replication with two genres of elementary school students' text—narrative and instructive—using nine LLMs from three providers (OpenAI, Anthropic, Mistral). Each LLM generated three independent scores per text via structured, benchmark-aligned prompts that were then aggregated into a consensus score. Results showed that intrarater reliability was high to excellent, ICC(3, *k*) ≈ .68–.97, and alignment with human ratings ranged from moderate to strong, ICC(3, 1) ≈ .47–.85, with larger models consistently outperforming smaller ones. Systematic bias patterns emerged, varying by model and genre, indicating a need for calibration. Increasing output token windows and reducing temperature parameters mitigated truncation and schema-related failures. Although LLM-based benchmark ratings can approximate expert judgments and reduce the need for labor-intensive human triple coding, limitations remain regarding cost (for larger models), genre- and task specificity, and sensitivity to text presentation and student grade level—factors that constrain immediate classroom use, particularly for formative feedback.

Keywords: text rating • large language models • narrative texts • instructive texts • elementary school

Didaktik Deutsch
Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur
Special Issue (2026) – Writing with AI. S. 42–64
DOI: 10.21248/dideu.963

1 Introduction

The effectiveness of feedback in writing instruction is well documented (Graham et al., 2011a; Scherer et al., 2024). Moreover, in both summative and formative feedback contexts, large language models (LLMs) hold considerable promise for supporting text rating.[1] Several arguments support their use:

- Text rating, whether for formative or summative purposes, is time-intensive. LLMs could help reduce teacher workload (Peltzer et al., 2025).
- Teacher-generated ratings often vary considerably (Birkel & Birkel, 2002). In contrast, LLMs generate responses based on probabilistic, distributional predictions (Zhao et al., 2023), potentially yielding less situationally influenced and more consistent assessments. By reducing susceptibility to subjective bias, LLM-based scoring may offer greater objectivity and rating quality.
- Process-oriented writing instruction and feedback on text revision—whether from teachers or peers—remain rare in classroom practice (Müller et al., 2023). Surveys among teachers and observations show that students predominantly receive summative feedback (Applebee & Langer, 2011; Roos, 2017) that focuses on final products rather than supporting students' writing during the writing process (Busse et al., 2022). This imbalance is probably related to the high demands on teacher time and effort. LLM-based feedback may offer a scalable way to provide timely, revision-oriented feedback that is integrated into the writing process.

To fully harness the potential of LLM-based feedback, however, LLM-based rating must meet the same requirements for objectivity, reliability, and validity as traditional text ratings. Otherwise, scores from LLM-based ratings will be too elastic or not suitable for teachers to make sound decisions about students' writing and their need to make progress as writers.

Following a brief overview of the current state of research, we present a study examining whether an established benchmark rating procedure can be replicated through structured prompting of multiple LLMs. The resulting ratings are compared with human ratings and evaluated using psychometric criteria (e.g., ICCs and mean-difference tests). Results indicate that LLM-based benchmark scoring is technically feasible following prompt and pipeline refinements, with virtually no unrecoverable failures. Across repeated runs, the models demonstrated high to excellent within-model consistency and moderate to strong overall alignment with human ratings. However, notable systematic biases emerged, varying by model and text genre—particularly for narrative texts and smaller models.

2 Empirical background

A recent meta-analysis on algorithm-based writing feedback found no statistically significant effects on either surface-level or deep-level writing outcomes. However, this is partly due to the limited number of eligible studies. Notably, this meta-analysis did not include studies on LLM-based feedback (Scherer et al., 2024). Early studies on the text-rating capabilities of LLMs present mixed results, with a slight overall trend toward high-quality text-rating performance (Altamimi, 2023; Bui & Barrot, 2024; Evmenova et al., 2024; Lu et al., 2024; Parker et al., 2023; Seßler et al., 2025; Shabara et al., 2024;

[1] In general, feedback and particularly feedback on texts also encompasses the evaluation of texts and the assessment of text quality. As used here, the term "feedback" subsumes summative and formative feedback as well as the assessment and rating of texts.

Steiss et al., 2024). However, the quality of these ratings has been shown to depend strongly on the structure and specificity of the prompt input (Jacobsen & Weber, 2025; Liu et al., 2023).

The existing studies address different aspects that can be categorized into distinct thematic areas. From these categories, several research gaps emerge that have yet to be explored systematically: consistency between LLM-based and human ratings; type of assessment; sensitivity to student characteristics, genres, and grade level; and single versus multiple LLMs.

(a) Consistency between LLM-based and human ratings

From both scientific and educational practice perspectives, it is highly relevant to determine the extent to which LLM-based ratings align with human ratings that themselves meet established standards of scoring validity and reliability. To understand the potential of LLM-based feedback, it is necessary to examine its performance through two lenses: the internal consistency of repeated iterations and the degree of correspondence with validated human ratings.

Seßler et al. (2025) report consistently good reliability of repeated evaluations using GPT-3.5, GPT-4, and GPT-o1, whereas LLaMA-3-70B and Mixtral-8×7B showed weaker internal agreement. However, only GPT-o1 achieved high correlations between LLMs and human raters. More variability was present when it came to content-related (deep-level) features, prompting the authors to suggest that LLMs may be more suitable for feedback on surface-level features. Whereas Scherer et al. (2024) and Seßler et al. (2025) assume, based on single studies, that algorithmic or LLM-based feedback is particularly accurate for assessing surface-level aspects, other researchers (e.g., Yan et al., 2024) argue that LLM-based text rating is also suitable for assessing deep-level features. However, it is important to note that interrater reliability among human raters is not always reported (e.g., Seßler et al., 2025), and this complicates interpretation of LLM–human alignment.

(b) Type of assessment

According to Schipolowski and Böhme (2016) assessments can follow different approaches: analytic (separate evaluation of various characteristics using criteria), holistic (overall evaluation of the text as a whole), or semiholistic (evaluation of distinct dimensions, but not fully criterion-separated). LLMs, however, do not analyze texts in a strictly criterion-based analytic manner—that is, they do not reliably decouple their evaluations along distinct, independently weighted criteria (Murugadoss et al., 2025). As a result, LLM-based ratings, particularly when prompted with multiple criteria, are best described as semiholistic. Even when individual criteria are specified in the prompt, the resulting rating is not analytical.

A new alternative is comparative assessment, which includes the benchmark rating approach (Bouwer et al., 2023, p. 305). In benchmark rating, texts are holistically positioned on a rating scale anchored by predefined benchmark texts. Each benchmark text exemplifies a specific quality level and is, for that purpose, enriched with textual descriptors. Empirical studies have shown that benchmark ratings meet the central psychometric quality standards of objectivity, reliability, and validity (Bouwer et al., 2023). To date, comparative or benchmark ratings using LLMs remain a desideratum. Given that benchmark ratings have been shown to be valid, reliable, and objective—and due to their clear operationalization that aligns well with structured LLM prompting—such procedures can be assumed to have high potential to be adapted for LLM ratings.

(c) Sensitivity to student characteristics

Previous studies have largely relied on relatively simple prompting strategies, occasionally including grade level and other student characteristics (e.g., second-language [L2] learners or struggling writers). This raises another desideratum: Evmenova et al. (2024) showed that LLM-based feedback does not always take information about students' abilities or needs into account, even though this information was provided in the prompt. It should be noted that Evmenova et al. (2024) used ChatGPT 3.5 and 4, and presumably also the smaller version of the model. With larger models, potentially better results could be achieved.

(d) Genre sensitivity

The studies cited above concentrated on one single text genre. Given that LLMs, as probabilistic language models, encode latent genre knowledge (Bubenhofer, 2025), it is essential to evaluate whether their rating performance remains consistent across different genres. However, prior research has largely been limited to single genres in isolation, most commonly focusing on persuasive opinion-based essays, argumentative texts, or source-based writing (Bui & Barrot, 2024; Evmenova et al., 2024; Lu et al., 2024; Steiss et al., 2024); and sometimes also narrative texts (Seßler et al., 2025).

(e) Grade level

Most studies to date focused on postsecondary writing, leaving other educational levels underexplored. Notable exceptions include the studies by Evmenova et al. (2024) and Seßler et al. (2025) that examine school-level texts. As shown by Evmenova et al. (2024), LLM-based rating shows a limited capacity to differentiate between grade levels when assessing student texts. This issue is particularly evident in the fact that alignment of LLM-based with human ratings is lower when evaluating lower grade level texts.

(f) Single versus multiple LLMs

Variations in training data composition, model architecture, and scale suggest that different LLMs probably exhibit divergent performance in text assessment tasks. Recent findings by Seßler et al. (2025) indicate that larger and more sophisticated models—albeit with higher operational costs—generally tend to outperform smaller models. However, a systematic comparison of multiple LLMs within a unified benchmark rating framework is still lacking. Such comparative analyses are essential to determine which models are most suitable for educational use—particularly in terms of reliability, cost-efficiency, and genre or age-group sensitivity.

3 Research questions

To address the research gaps identified above, especially regarding (a) consistency, (b) type of assessment, (d) several genres, (e) lower grade levels and (f) multiple LLMs, the present study tests a benchmark-aligned prompting procedure across two text genres (narrative and instructive), lower grade levels (4 and 5), and multiple LLMs from different providers (see below). The study aims to answer the following research questions:

(1) Can a comparative rating procedure, such as benchmark rating, be successfully implemented into an LLM-based assessment system?

(2) Do LLM-generated benchmark ratings demonstrate sufficient intrarater reliability across repeated assessments, and is this reliability consistent across different model providers and architectures?

(3) To what extent do LLM-generated benchmark scores correlate with human expert ratings, and are there systematic differences in scoring patterns between human raters and LLMs?

These questions are central not only to understanding the general capability of LLMs for educational text evaluation but also to evaluating their potential for use in teaching and learning settings (e.g., model solutions in the classroom, automated formative feedback, and integration into digital learning environments).

4 Method

4.1 Sample

The present study—conducted as part of the DEEP consortium (https://deep-consortium.ch/de) and funded by the Jacobs Foundation—uses data from another project funded by the Swiss National Science Foundation.[2] This project was designed as an intervention study with three measurement points and aimed to investigate whether a writing-fluency training, implemented in addition to a process-oriented writing program, enhances students' writing competencies (Grade 4/5), and whether students with German as a second language (L2) benefit more than students with German as a first language (L1) (for further details on the intervention study, see Sturm & Schneider, 2021). The data used in the present study stem from all three measurement points.

To ensure transparency of the chosen approach, Section 4.2 provides a detailed outline of the benchmark rating procedure, and Section 4.3 outlines the design of the LLM-based system and the technical implementation, including information on the selected models.

4.2 Material and rating procedure

Students completed two writing tasks—one narrative and one instructive—each within a 30-min timeframe. The narrative writing task was based on a prompt adapted from Schneider et al. (2012): "It was a rainy Sunday. Family Racconti was still asleep."[3] Students were asked to continue the story and write an exciting narrative text. The instructive writing task required students to write step-by-step directions for drawing a specific picture. They were shown an image of a snowman standing on a meadow in the sun and instructed to trace the image so that someone unfamiliar with it could reproduce it based on the instructions.

To eliminate potential presentation effects (Graham et al., 2011b),[4] the handwritten texts were transcribed and corrected for spelling and grammar. The resulting dataset from all three measurements comprised 873 narrative texts (mean text length M = 820.52 characters, SD = 412.07 characters,

[2] "Fostering writing fluency," grant number 100019_159311

[3] The story prompt was given in German: "*Es war ein verregneter Sonntag. Familie Racconti schlief noch*." The story prompt was slightly modified at each measurement point: "There was thick fog outside. The Racconti family were having dinner." "It was snowing heavily outside. The Racconti family were having lunch." Despite these changes, the prompts remained comparable in structure and intent, each setting the stage for a narrative involving the Racconti family.

[4] Surface-level features such as spelling or grammar errors may influence raters' judgments about text quality.

Min = 62 characters, Max = 2,586 characters) and 859 instructive texts (mean text length *M* = 459.95 characters, *SD* = 243.35 characters, Min = 56 characters, Max = 1,502 characters).

The quality of the texts was assessed using a benchmark rating. To this end, a scale of five benchmark texts representing varying quality levels was developed following a five-step procedure outlined by Feenstra (2014). Steps 1–4 involved the random selection of 48 narrative and 48 instructive texts from the full sample.

(1) A group of three experienced raters formulated genre- and task-specific criteria deemed essential for assessing text quality. These criteria were aligned with widely accepted dimensions of writing assessment, content, structure, and language; and they tailored the two task types: For the narrative task, criteria focused on the central idea of the story, the narrative structure, and the use of linguistic devices to create tension or to differentiate characters or other narrative elements. For the instructive task, criteria included the reader-oriented clarity of the instructions, the descriptive quality of the image, the logical sequencing of the instructional steps, and the use of cohesive linguistic markers.

(2) Based on these criteria from Step 1, the raters selected a representative "average" benchmark text from the sample that demonstrated both strengths and weaknesses across the defined dimensions. This text was assigned an arbitrary score of 100, accompanied by a qualitative rationale that detailed why the text was considered of average quality with reference to the task-specific criteria.

(3) The raters from Step 1 then scored the remaining randomly selected student texts by comparing each one directly to the average benchmark text. Each text received a numerical score indicating how it compared to the benchmark: Thus, if a text seemed only half as good as the benchmark text, it received a score of 50; if it seemed twice as good, it was assigned a score of 200. Interrater reliability for this comparative prerating was good[5] (narrative: .89 for ICC[3, 1]; instructive: .86 for ICC[3, 1]). To control for individual rater leniency and severity, all scores were *z*-standardized. Next, an average score was computed per text across all three judgments. These means were then transformed to a new scale with a mean of 100 and a standard deviation of 15 to facilitate the construction of a benchmark scale.

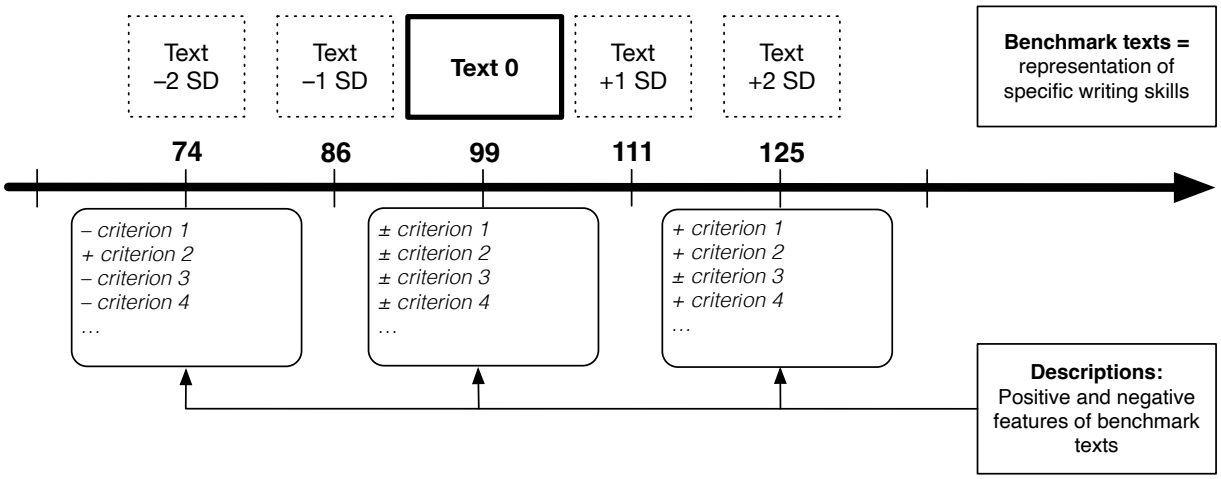
(4) Based on the transformed scores, five texts per genre were selected to represent different levels of writing quality with (mean) ratings around 70, 85, 100, 115, and 130. Each of the five selected benchmark texts was supplemented with a descriptor summarizing its key strengths and weaknesses according to the established criteria. Figure 1 illustrates this for the narrative benchmark texts (see also Appendix A with the narrative benchmark Text 0 and accompanying descriptions).

(5) Using the newly constructed benchmark scale, an independent group of three trained raters (different from those in Step 3) rated all texts from the three measurement points in randomized order. Each text was systematically compared to the benchmark texts and assigned a score that places it along the scale. The text under evaluation was first compared with Text 0; if it differed from Text 0, it was then compared with the next benchmark text at +1 *SD* or –1 *SD*, and, if necessary, at +2 *SD* or –2 *SD*.

[5] Koo and Li (2016) classify values above .75 as "good". In the context of complex educational writing assessments, such coefficients are often regarded as sufficient for consequential use of ratings (Graham et al., 2012).

Figure 1

Schematic Representation of the Five Narrative Benchmark Texts with Their Respective Scores



Most raters involved in Step 5 were student teachers, although a smaller number were research assistants participating in Sturm and Schneider's (2021) intervention study. All raters received a half-day training session supported by a detailed rating manual. During training, raters practiced on a selection of texts of varying quality and discussed their judgments with the trainer to align on the benchmark-based rating approach. The three raters then independently scored a set of 50 additional texts. Provided that interrater reliability reached an acceptable level, they proceeded to rate the remaining texts independently. For the instructive texts, a second training was required due to uncertainties in handling student texts that were purely descriptive or narrative, thereby deviating from the task expectations. Interrater reliability of the final human ratings was estimated using the intraclass correlation, yielding an excellent reliability of ICC(3, 1) = .99. For all subsequent analyses, the mean score across the three raters was computed for each text.

4.3 Design of the LLM-based system

To translate the benchmark rating procedure into an LLM-based system, we designed genre-specific system prompts and individualized user prompts for each student text. The system prompt employed several evidence-based prompt engineering strategies to enhance model performance and reliability.

Specifically, the system prompt was structured to incorporate multiple complementary strategies. First, it used human-aligned instruction design by mirroring both the benchmark descriptors and the training manual used to calibrate human raters, thereby anchoring the LLM to established expert standards as described in Section 4.2. Second, it implemented a step-by-step reasoning specification via an explicit, hierarchical outline of the assessment criteria, effectively using a chain-of-thought (CoT) approach.[6] Third, the prompt incorporated explicit reasoning extraction, instructing the model to articulate and return its rationale for each score. Finally, the system prompt employed structured output formatting, requiring all benchmark scores to be in a machine-readable JSON format, facilitating

[6] Chain-of-thought prompting is a technique that enables LLMs to solve complex tasks by decomposing them into a series of intermediate reasoning steps. This process mimics human-like problem solving, improving the model's accuracy and transparency in decision making (Wei et al., 2023).

downstream processing, and minimizing ambiguity in score extraction. The full system prompt for scoring narrative texts is provided in Appendix B.

Rather than submitting requests synchronously, we employed batch processing[7] across multiple LLM providers via each provider's native batch API. This asynchronous, provider-specific batch handling reduced costs while enabling high-throughput processing. For each model–corpus combination, we generated a provider-specific JSONL batch file[8] containing the system prompt, the instantiated user prompt, and structured schema constraints. The user prompt template for the narrative text scoring is shown in Appendix C. Responses that failed schema validation or resulted in API errors were logged separately, along with corresponding error messages, allowing for subsequent error analysis.

To answer the first research question, we implemented this proof of concept with three LLM providers: OpenAI and Anthropic as well-established commercial providers, and Mistral, representing a European-based provider. In total, we evaluated student texts across nine distinct LLM variants, spanning these three providers and representing different performance levels (see Table 1). Notably, we did not use reasoning-augmented models in this study, because these entail substantially higher computational costs than standard language models, and our study prioritizes cost-effective applications suitable for research and educational contexts. The step-by-step reasoning specification we provided in the system prompt already enables LLMs to articulate abbreviated reasoning processes and justifications for text assessment.

Table 1

Providers and Models of Different Performance Levels

Provider	Large	Medium	Small
Mistral	mistral-medium-2505	open-mistral-nemo-2407	ministral-8b-2410
Anthropic	claude-sonnet-4-20250514	claude-3-5-haiku-20241022	claude-3-haiku-20240307
OpenAI	gpt-4.1-2025-04-14	gpt-4.1-mini-2025-04-14	gpt-4.1-nano-2025-04-14

All requests were executed using standardized generation parameters. The maximum token[9] output was set to 1,024 tokens, and the temperature[10] was set to standard values used in the chat interface for end users, enabling consistent behavior across evaluations. The temperature parameter controls the randomness of token selection by scaling the probability distribution of output tokens, thereby affecting the determinism of model responses (Li et al., 2025).

To answer the second research question, we refined both the prompt design and processing pipeline based on insights from the initial proof-of-concept implementation and identified failure modes. While

[7] Unlike synchronous processing (real-time interaction), batch processing allows a large number of requests to be sent simultaneously and processed by the provider at a lower priority. This reduces costs.

[8] JSONL (JSON Lines) is a text format in which each line is a valid JSON object, allowing for efficient processing of large datasets because each record can be handled independently.

[9] A token is the basic unit of text processed by an LLM, representing a sequence of characters, a word, or part of a word.

[10] The temperature parameter ranges typically from 0 to 1 (or 2). Lower values (e.g., < 0.3) make the model more deterministic and focused, whereas higher values increase creativity and diversity but also the risk of "hallucinations" or inconsistencies.

retaining the same set of LLM providers and models, we enhanced the system prompt to elicit more explicit comparisons between student texts and benchmark exemplars. We increased the maximum token output to 2,048 tokens to prevent truncation, and the temperature was fixed at 0.3, allowing for more deterministic responses that align with the structured output schema.[11]

To assess the intrarater reliability of LLM scoring, each student text was evaluated three independent times by each of the nine models. We then computed ICCs using a two-way mixed effects model for mean-rating consistency—that is, ICC(3, *k*) with *k* = 3 raters.

To address the third research question concerning the validity of automated LLM-based scoring, we compared LLM-generated scores to human benchmark ratings. For each text, an aggregate LLM score was calculated by averaging the results from the three independent scoring runs. To handle missing data, we applied a systematic protocol: A missing score in a single evaluation run was excluded from that run's contribution, and the aggregate score was calculated from the remaining two evaluations. Texts with two or more missing scores across the three runs were excluded from further analysis due to insufficient evaluation data. These aggregate LLM scores were then compared to the human expert benchmark scores using a two-way mixed-effects model for consistency of a single rater—that is, ICC(3, 1). To examine systematic differences between human and LLM scoring, we computed the mean difference between LLM and human scores for each text, along with corresponding standard deviations. Paired-samples *t*-tests were conducted to test for significant mean differences at the aggregate level and Cohen's *d* was reported as a measure of effect size to characterize the magnitude of observed differences.

5 Results

First, we report findings from the initial implementation of our LLM-based benchmark scoring system across nine LLM variants (three providers, three models per provider, see Table 1). The aim was to assess feasibility and identify any systematic failure pattern in schema validation and output generation. Second, we present reliability analyses based on three independent scoring runs per text, quantifying intrarater consistency using ICC(3, *k*) with 95% confidence intervals (CI). Third, we evaluate criterion validity by comparing aggregate LLM-generated scores against human expert benchmark ratings using ICC(3, 1) and paired-samples *t*-tests, with effect sizes (Cohen's *d*) characterizing the magnitude of any systematic differences.

The LLM-based benchmark rating system was shown to be technically feasible across all nine model variants, despite notable heterogeneity in initial performance. During the full proof-of-concept phase, the system successfully generated scoreable outputs for the majority of student texts. Based on the error analysis from the proof-of-concept phase, two critical issues emerged that required methodological adjustments in the subsequent phase: (a) The primary failure mode across all LLM providers was output truncation at the 1,024-token limit. In 89.81% of these cases, extraction errors could be resolved through post hoc automated recovery. Nonetheless, the underlying root cause was systematic: Models tended to generate reasoning justifications that exceeded the token budget. (b) The error patterns also revealed that higher temperature (default chat settings) contributed to verbose,

[11] Although lower temperature values are known to reduce output randomness and promote more deterministic model behavior, specific temperature recommendations in the literature are limited. The value of 0.3 was selected based on common practice in applications requiring consistent, structured output, though we note this reflects practitioner guidance rather than empirically derived thresholds from academic research.

unfocused reasoning, which often failed to align with the structured output schema. To address these issues, two adjustments were implemented: The maximum token limit was increased to 2,048 tokens, and the temperature was reduced to 0.3 to promote more deterministic responses. In the refined scoring phase, the system yielded just 0.02% unrecoverable failures, which were treated as missing values in subsequent analyses.

To address our second research question, we evaluated the consistency and reliability of LLM-based scoring across three repeated independent evaluation runs (see Table 2). Following Koo and Li (2016), results revealed moderate to excellent intrarater reliability across all nine LLM variants and both text genres.

- Narrative texts: ICC(3, *k*) values ranged from .77 to .95 (*M* = .88), falling within the good to excellent range. The most consistent models were OpenAI GPT-4.1 and Claude Sonnet.
- Instructive texts: Reliability was comparable or higher, with ICC(3, *k*) values between .68 and .97 (*M* = .87), again with OpenAI GPT-4.1 achieving the highest reliability. Smaller models fell into the moderate category.

These findings confirm that the refined prompt design combined with increased token budget and reduced temperature successfully enhanced the determinism and reproducibility of the scoring system.

Table 2

Intraclass Correlation of Three Individual Scores Produced by the Same LLM by the Benchmark Rating Procedure

Corpus	Model	Intraclass correlation ICC(3, *k*) and 95% CI		
		Estimate	Lower limit	Upper limit
Instructive	gpt-4.1-2025-04-14	.973	.969	.976
	claude-sonnet-4-20250514	.955	.949	.960
	claude-3-5-haiku-20241022	.949	.943	.955
	mistral-medium-2505	.942	.935	.949
	gpt-4.1-mini-2025-04-14	.936	.928	.943
	claude-3-haiku-20240307	.877	.862	.891
	open-mistral-nemo-2407	.765	.736	.791
	ministral-8b-2410	.744	.712	.773
	gpt-4.1-nano-2025-04-14	.684	.644	.720
Narrative	claude-sonnet-4-20250514	.952	.947	.958
	gpt-4.1-2025-04-14	.952	.946	.957
	claude-3-5-haiku-20241022	.924	.915	.932
	gpt-4.1-mini-2025-04-14	.907	.896	.917
	mistral-medium-2505	.896	.883	.907
	claude-3-haiku-20240307	.876	.861	.890
	open-mistral-nemo-2407	.859	.842	.875
	gpt-4.1-nano-2025-04-14	.772	.744	.797
	ministral-8b-2410	.766	.738	.792

Note. For each corpus the models are sorted by the point estimate of the ICC(3, *k*) value in descending order.

To address our third research question, we compared aggregate LLM scores (like in human-based rating—cf. Section 4.2—, averaged across three independent runs) with human expert benchmark ratings (see Table 3). In 14 cases, the aggregate LLM score was based on only two valid runs due to missing values. For narrative texts, the alignment of LLM-generated with human ratings ranged from poor to good: ICC(3, 1) = .47 to .85 *(M* = .71). OpenAI GPT-4.1 demonstrated the highest alignment with human benchmarks, followed by OpenAI GPT-4.1-mini and Claude Sonnet. In contrast, the smaller Mistral Ministral-8b model showed substantially lower alignment. For instructive texts, alignment was similarly variable: ICC(3, 1) = .61 to .85 (*M* = .75), ranging from moderate to good. Mistral Medium, Claude Sonnet, and OpenAI GPT-4.1 achieved the highest validity for instructive texts, whereas smaller models again showed reduced alignment with human ratings.

Table 3

Intraclass Correlation Between Human and LLM-Based Mean Scores Produced by the Benchmark Rating Procedure

Corpus	Model	Intraclass correlation ICC(3) and 95% CI		
		Estimate	Lower limit	Upper limit
Instructive	mistral-medium-2505	.848	.828	.865
	claude-sonnet-4-20250514	.830	.808	.849
	gpt-4.1-2025-04-14	.819	.796	.840
	gpt-4.1-mini-2025-04-14	.801	.776	.824
	claude-3-haiku-20240307	.793	.767	.817
	claude-3-5-haiku-20241022	.725	.691	.755
	ministral-8b-2410	.683	.646	.718
	open-mistral-nemo-2407	.618	.575	.658
	gpt-4.1-nano-2025-04-14	.608	.564	.649
Narrative	gpt-4.1-2025-04-14	.854	.835	.871
	gpt-4.1-mini-2025-04-14	.808	.784	.830
	claude-sonnet-4-20250514	.787	.760	.811
	mistral-medium-2505	.770	.741	.796
	gpt-4.1-nano-2025-04-14	.712	.678	.743
	claude-3-5-haiku-20241022	.682	.645	.716
	open-mistral-nemo-2407	.638	.597	.676
	claude-3-haiku-20240307	.637	.595	.674
	ministral-8b-2410	.471	.418	.521

Note. For each corpus the models are sorted by the point estimate of the ICC(3) value in descending order.

Paired-samples *t*-tests and Cohen's *d* effect sizes revealed substantial systematic bias patterns across the evaluated models (see Table 4). For instructive texts, four models, with small to medium performance levels, showed negligible differences from human ratings. Claude Sonnet showed a small underestimation of text quality compared to the human ratings, whereas OpenAI GPT-4.1 showed a large overestimation. In contrast, the other small models, OpenAI GPT-4.1-nano and Mistral Ministral, showed substantial systematic underestimation of text quality.

Table 4

Mean Difference Analysis of Human and LLM-Based Mean Scores Produced by the Benchmark Rating Procedure

Corpus	Model	Difference between human and LLM-based mean scores		Paired sample *t*-test			Cohen's *d*
		M	*SD*	*t*	*df*	*p*	
Instructive	gpt-4.1-nano-2025-04-14	18.976	9.774	54.913	799	< .001	1.941
	ministral-8b-2410	11.589	8.205	39.825	794	< .001	1.412
	claude-sonnet-4-20250514	1.290	6.918	5.355	824	< .001	0.186
	claude-3-haiku-20240307	0.600	7.340	2.300	790	.022	0.082
	claude-3-5-haiku-20241022	−0.219	9.855	−0.630	806	.529	−0.022
	gpt-4.1-mini-2025-04-14	−0.194	8.274	−0.671	822	.502	−0.023
	open-mistral-nemo-2407	−0.457	11.587	−1.133	824	.258	−0.039
	mistral-medium-2505	−3.276	6.306	−14.920	824	< .001	−0.519
	gpt-4.1-2025-04-14	−5.562	7.754	−20.614	825	< .001	−0.717
Narrative	ministral-8b-2410	17.642	10.596	48.940	863	< .001	1.665
	gpt-4.1-nano-2025-04-14	8.348	9.633	25.472	863	< .001	0.867
	claude-3-5-haiku-20241022	6.473	10.183	18.683	863	< .001	0.636
	gpt-4.1-mini-2025-04-14	5.603	9.157	17.988	863	< .001	0.612
	open-mistral-nemo-2407	5.299	11.035	14.114	863	< .001	0.480
	claude-3-haiku-20240307	5.011	10.574	13.922	862	< .001	0.474
	gpt-4.1-2025-04-14	2.068	7.436	8.176	863	< .001	0.278
	claude-sonnet-4-20250514	2.248	10.230	6.458	863	< .001	0.220
	mistral-medium-2505	−1.862	8.943	−6.119	863	< .001	−0.208

Note. A positive value for the mean difference between human and LLM-based mean scores indicates that humans assigned higher scores to the student texts than the LLM-based scoring system. For each corpus, models are sorted by the effect size in descending order.

For narrative texts, systematic bias was more pronounced. All models showed significant differences to those of human ratings. The majority of models significantly underestimated text quality. Mistral Ministral showed the largest bias followed by OpenAI GPT-4.1 nano—both representing large effect sizes. Other models also substantially overestimated text quality, though with small to medium effects. The closest alignment with human judgments was found in Claude Sonnet, Open AI GPT-4.1, and Mistral Medium—all showing small but statistically significant differences. Notably, Mistral Medium was the only model to systematically overestimate the quality of narrative texts.

6 Discussion

This study examined whether an established human rating procedure (benchmark rating) can be replicated using LLMs, and whether such replication meets key psychometric criteria. Specifically, the latter question must be resolved affirmatively before an LLM-based assessment of this kind can be authorized for implementation in schools. Overall, findings indicate that LLM-based benchmark rating is feasible, demonstrates good alignment with human ratings, and shows acceptable reliability and validity. However, several challenges remain before LLM-based assessment can be reliably integrated into educational practice.

In response to our first research question, results confirm that LLM-based benchmark rating is operationally viable across a diverse set of models and providers. However, successful implementation required prompt refinement and parameter optimization—specifically, increased output windows and reduced temperature settings—to reduce schema failures and ensure consistency. Without such technical adjustments, initial failure rates were substantial.

When texts are rated through comparative benchmark rating, there is generally less room for subjective interpretation than in analytical or (semi)holistic approaches (Lindauer & Sommer, 2018). Consequently, interrater reliability among human raters tends to be very high. This makes comparative benchmark rating highly relevant for educational use. Regarding our second research question, findings show that the consistency of LLM-based ratings across all nine LLM variants and both text genres was satisfactory to excellent. As expected, larger, more sophisticated models (OpenAI GPT-4.1, Claude Sonnet) outperformed models of medium and small performance levels and demonstrated higher consistency.

Regarding our third research question, LLM-generated scores were found to correlate meaningfully with human expert ratings across models and genres. However, systematic differences were also observed: Larger models demonstrated closer alignment with human standards and smaller effect sizes in deviation from human scores. The smallest models (OpenAI GPT-4.1 nano and Mistral ministral-8b variants) showed larger, consistent biases. Furthermore, genre-specific patterns emerged: Instructive texts were often underestimated (or scored conservatively) by midsize and large models, whereas narrative texts tended to be overestimated, particularly by smaller models. These findings suggest that genre and model-specific calibration is essential when applying LLM-based assessments in practice.

It should be noted that the LLMs used in this study—provided by OpenAI, Anthropic, and Mistral—are trained on broad but nonspecific datasets. As the SWK (2023) emphasizes, transparency regarding training data is critical in educational contexts. The report also highlights the potential of "model solutions" to improve LLM-based feedback. In our study, we operationalize this concept by using comparisons with benchmark texts as a model solution. These serve as transparent reference standards that communicate task- and level-specific performance expectations to the LLM. However, a model solution is more than a tool for achieving high statistical alignment of human with LLM-based ratings. It acts as an anchor that makes performance expectations explicit and verifiable. This shifts the evaluative focus from correlational agreement to alignment with educational standards. In our study, the benchmark texts represent a specific stage in the learning trajectory, namely from Midgrade 4 to Midgrade 5. It follows that different grade levels or genres, such as argumentative writing, would require their own distinct benchmark texts to ensure the model solution accurately reflects appropriate expectations.

Whereas previous studies have focused largely on academic or argumentative texts (e.g., persuasive opinion essays or source-based writing), our findings demonstrate that LLM-based benchmark ratings can be successfully applied to both narrative and instructive texts. Notably, results yield slightly better validity for instructive texts than for narrative texts (e.g., higher average alignment with human ratings, ICC(3, 1) with *M* = .75 and range .61–.85 versus *M* = .71 and range .47–.85). This probably reflects that the instructive task is more tightly specified: The writing goal and audience are explicit, and the task permits less variation in acceptable solutions than a narrative task. In contrast, the narrative prompt is more open-ended, leading to highly diverse responses: Some students wrote adventure stories; others fantasy, crime, or even a drama with an unhappy ending. Human raters appear somewhat better equipped to handle such open-ended tasks. To enable broad implementation of LLM-based benchmark rating across genres in educational settings, further research with additional genres is required.

It remains unclear to what extent the presentation effect arises in LLM-based feedback in a manner analogous to human rating, or whether it can be substantially reduced. This question would arise particularly if—unlike in the present study—uncorrected texts were to be assessed using an LLM-based approach. As Sturm et al. (2024) show, presentation effects can also be demonstrated in human-based benchmark rating. When we use the same texts in their uncorrected form, we find initial indications that LLM-based feedback is likewise subject to this effect (Sturm et al., 2026).

The various providers tested are comparable in terms of assessment quality. However, larger models both perform better and appear less susceptible to distortions. This complicates the use of LLM-based ratings in school settings, because these models are substantially more expensive than smaller LLMs. Based on these findings, efforts by some German federal states to develop their own LLMs should be viewed with caution, because such models are unlikely to match the capabilities of the large international systems. If schools nevertheless opt to use low-cost or unverified models, it can be expected that students' writing performance will not be assessed reliably, and that instructional decisions based on such assessments may likewise be inappropriate.

Evidence from pre-LLM-automated writing evaluation indicates that such feedback is most effective when part of a larger instructional program in which, for example, teachers provide strategy instruction, iterative practice, or individual support (Nunes et al., 2022). Although evidence is lacking for LLM-based feedback, it is reasonable to assume that these pedagogical principles are transferable to this new generation of feedback tools.

The development of comprehensive instructional programs and materials that integrate the potential of LLMs therefore appears relevant. Although LLMs can replicate text assessment, the ultimate responsibility for assessing texts should remain with teachers (SWK, 2023). Depending on which LLM-based tools are available and actually used—whether lower-cost or higher-cost, validated for the respective grade level, and validated for various genres—teachers would need to pay closer attention particularly to students with lower writing performance.

Limitations

As mentioned above, a key limitation of the current study is that only two text types have been evaluated so far. Future work should extend the approach to additional genres. A further limitation, particularly complicating use in educational contexts, is that benchmark ratings have thus far been restricted to the closed setting of specific writing tasks for which benchmark texts are available. Developing such benchmarks remains a manual, time-intensive process. Furthermore, our approach currently provides only summative feedback. Generating formative feedback—which would ultimately provide real relief

for teachers—still requires substantial development. This would require linking total scores to pedagogically useful descriptors that identify text-specific strengths and areas for improvement in a learner-appropriate way. Benchmark ratings appear promising in this regard, because they are inherently descriptor-based and align with defined performance expectations. Initial attempts in this area are already promising. In addition to intended use in teaching and learning settings, it is already conceivable that LLM-based benchmark rating could also serve as a testing tool in research projects—for instance, in cross-sectional assessments of students' writing skills, similar to DESI (Neumann & Lehmann, 2008) or NAEP (2017).

References

Altamimi, A. B. (2023). Effectiveness of ChatGPT in essay autograding. In *2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (pp. 102–106). https://doi.org/10.1109/iCCECE59400.2023.10238541

Applebee, A. N., & Langer, J. A. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal*, 100(6), 14–27. https://doi.org/10.58680/ej201116413

Birkel, C., & Birkel, P. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, *49*, 219–224.

Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., & De Maeyer, S. (2023). Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement. *Journal of Writing Research*, *15*(3), 497–518. https://doi.org/10.17239/jowr-2024.15.03.03

Bubenhofer, N. (2025). Vektorisierte Praktiken: Ein linguistisches Verständnis von Large Language Models. *Leseräume*, *11*, 1–16.

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Busse, V., Müller, N., & Siekmann, L. (2022). Wirksame Schreibförderung durch diversitätssensibles formatives Feedback. In V. Busse, N. Müller, & L. Siekmann (Eds.), *Schreiben fachübergreifend fördern. Theoretische Grundlagen und Praxisanregungen für Schule, Unterricht und Lehrerinnen- und Lehrerbildung* (pp. 114–133). Klett Kallmeyer.

Evmenova, A. S., Regan, K., Mergen, R., & Hrisseh, R. (2024). Improving writing feedback for struggling writers: Generative AI to the rescue? *TechTrends, 68*(4), 790–802. https://doi.org/10.1007/s11528-024-00965-y

Feenstra, H. (2014). *Assessing writing ability in primary education: On the evaluation of text quality and text complexity* (Doctoral dissertation). Universiteit Twente.

Graham, S., Harris, K. R., & Hebert, M. (2011a). *Informing writing: The benefits of formative assessment*. Carnegie Corporation of New York.

Graham, S., Harris, K. R., & Hebert, M. (2011b). It is more than just the message: Presentation effects in scoring writing. *Focus on Exceptional Children, 44*(4), 1–12.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform. https://files.eric.ed.gov/fulltext/ED532068.pdf

Jacobsen, L. J., & Weber, K. E. (2025). The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback. *AI, 6*(35), 1–17. https://doi.org/10.3390/ai6020035

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Li, L., Sleem, L., Gentile, N., Nichil, G., & State, R. (2025). Exploring the impact of temperature on large language models: Hot or cold? *arXiv*. https://doi.org/10.48550/arXiv.2506.07295

Lindauer, N., & Sommer, T. (2018). Verfahren der Textbeurteilung: Merkmale und Vorzüge eines holistischen Benchmarkratings. *Leseräume*, 5, 1–14.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), Article 195, 1–35. https://doi.org/10.1145/3560815

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education, 49*(5), 616–633. https://doi.org/10.1080/02602938.2024.2301722

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

Murugadoss, B., Poelitz, C., Drosos, I., Le, V., McKenna, N., Negreanu, C. S., Parnin, C., & Sarkar, A. (2025). Evaluating the evaluator: Measuring LLMs' adherence to task evaluation instructions. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(18), 19589–19597. https://doi.org/10.1609/aaai.v39i18.34157

NAEP. (2017). *Writing framework for the 2017 National Assessment of Educational Progress*. National Assessment Governing Board.

Neumann, A., & Lehmann, R. H. (2008). Schreiben Deutsch. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* (pp. 89–103). Beltz.

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning, 38*(2), 599–620. https://doi.org/10.1111/jcal.12635

Parker, J. L., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721–727. https://doi.org/10.3928/01484834-20231006-02

Peltzer, K., Lira Lorca, A., Krause, U.-M., Graham, S., Panadero, E., & Busse, V. (2025). How to support at-risk writers: Differential effects of formative feedback on argumentative writing and motivation. *Reading and Writing*, *39*, 965–1000. https://doi.org/10.1007/s11145-025-10652-w

Roos, M. (2017). *QUIMS-Bericht: Entwicklungen von 2014 bis 2016 (provisorische Fassung)*. spectrum GmbH.

Scherer, S., Graham, S., & Busse, V. (2024). How effective is feedback for L1, L2, and FL learners' writing? A meta-analysis. *Learning and Instruction, 93*, Article 101961. https://doi.org/10.1016/j.learninstruc.2024.101961

Schipolowski, S., & Böhme, K. (2016). Assessment of writing ability in secondary education: Comparison of analytic and holistic scoring systems for use in large-scale assessments. *L1-Educational Studies in Language and Literature, 16*, 1–22. https://doi.org/10.17239/L1ESLL-2016.16.01.03

Schneider, H., Wiesner, E., Lindauer, T., & Furger, J. (2012). Kinder schreiben auf einer Internetplattform: Resultate aus der Interventionsstudie myMoment2. *dieS-online, 2*, 1–37.

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 *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 462–472). ACM. https://doi.org/10.1145/3706468.3706527

Shabara, R., El Ebyary, K., & Boraie, D. (2024). Teacher or ChatGPT: The issue of accuracy and consistency in L2 assessment. *Teaching English with Technology, 24*(2), 71–92.

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 on students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Sturm, A., Lindauer, N., & Neuenschwander, M. P. (2024, June 27). *Benchmark ratings and presentation effects* [presentation]. SIG Writing, Paris, France.

Sturm, A., & Schneider, H. (2021). Flüssiges Formulieren in der Textproduktion (Klasse 4/5). *Didaktik Deutsch*, *26*(51), 28–49.

Sturm, A., Unger, V., & Grünig, F. (2026). LLM-based rating of texts and presentation effects. (Manuscript in preparation). Pädagogische Hochschule FHNW, Windisch, Switzerland.

SWK (Ständige Wissenschaftliche Kommission der Kultusministerkonferenz). (2023). *Large Language Models und ihre Potenziale im Bildungssystem: Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz*. SWK. https://doi.org/10.25656/01:28303

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024, September 5). Promises and challenges of generative artificial intelligence for human learning. *arXiv*, 1–36. https://doi.org/10.48550/arXiv.2408.12143

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., . . . Wen, J.-R. (2023). A survey of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2303.18223

Authors' addresses:

Afra Sturm, Pädagogische Hochschule FHNW, Bahnhofstrasse 6, 5210 Windisch, Switzerland
afra.sturm@fhnw.ch
Valentin Unger, Pädagogische Hochschule St. Gallen, Müller-Friedbergstrasse 34, 9400 Rorschach, Switzerland
valentin.unger@phsg.ch
Fabian Grünig, Pädagogische Hochschule St. Gallen, Müller-Friedbergstrasse 34, 9400 Rorschach, Switzerland
fabian.gruenig@phsg.ch

Appendix A

Narrative benchmark text (average text [Text 0], 99 points) – corrected and typed version	
Es war ein verregneter Sonntag. Familie Racconti schlief noch. Plötzlich wachen sie auf und hören ein sehr komisches Geräusch. Herr Racconti ist schauen gegangen, und vor seiner Tür war ein Hund. Herr Racconti hatte Angst vor dem Hund. Er ist einfach irgendwo anders gerannt und der Hund ist nach draussen gegangen. Herr Racconti hat so Angst gehabt, dass er eine halbe Stunde in seinem Zimmer versteckt war. Nach einer halben Stunde ist er gekommen, und er war sehr ruhig am Laufen. Dann hat er im ganzen Haus gesucht. Und er hat gesehen, dass der Hund nicht mehr dort war. Nach der Katastrophe ist er und seine Familie spazieren gegangen. Sie sind ins Schwimmbad gegangen. Der Sohn Yorby ist von einem 10-Meter-Sprungbrett gesprungen. Splash!! Er ist mit dem Bauch heruntergefallen. Sein Bauch war voll rot, und er hat nicht geweint. Er ist ein richtiger Jungmann. Die Woche sind sie ganz viel spazieren gegangen. Sie sind in den MacDonalds, ins Kino, einkaufen gegangen, weil er Euro-Millions gewonnen hat. Braaaa! Und er ist reich und froh für das ganze Leben geworden. Ende.	It was a rainy Sunday. The Racconti family were still asleep. Suddenly, they woke up to a strange noise. Mr. Racconti went to investigate and found a dog outside the front door. He was afraid of the dog. He ran off and the dog went outside. Mr. Racconti was so scared that he hid in his room for half an hour. After half an hour, he emerged and walked very calmly. Then he searched the whole house. He saw that the dog was no longer there. Afterwards, he and his family went for a walk. They went to the swimming pool. Their son Yorby jumped off the 10-metre diving board. Splash! He landed on his stomach. His belly turned red, but he didn't cry. He's a real young man. They went for lots of walks that week. They went to McDonald's, the movie theater, and the shops because he had won the EuroMillions lottery. Bravo! He became rich and happy for the rest of his life. The end.

Positive Aspects of the average text	
Inhalt 1) Im ersten Teil des Textes ist eine Ge samtidee erkennbar.) Es wird ansatzweise Spannung aufgebaut («… und hören ein sehr komisches Geräusch», «Herr Racconti hatte Angst»). 2) Der Text beinhaltet mehrere Figuren und mehrere Orte. Die Figur des Vaters wird im Verlauf etwas differenziert. Der Text baut ansatzweise eine narrative Handlungsstruktur auf (Vater rennt weg und schleicht durchs Haus → Bezug zur Anfangssituation). **Struktur** 3) Der narrative Aufbau ist teilweise vorhanden: ansatzweise Komplikation, eine Art Auflösung. 4) Innerhalb der einzelnen Teile des Textes (Hauptidee und weitere Ideen) ist die logische Abfolge vorhanden.	**Contents** 1) The main idea becomes apparent in the first part of the text. Some suspense is built up through phrases such as "and hear a very strange noise" and "Mr. Racconti was afraid." 2) The text features multiple characters and locations. The father's character becomes more nuanced as the story progresses. The text begins to establish a narrative structure (the father runs away and sneaks through the house, referencing the initial situation). **Structure** 3) The narrative structure is partially present, with a hint of complication and a kind of resolution. 4) There is a logical sequence within the individual parts of the text (main idea and other ideas).

Sprache 5) Es gibt vereinzelt Ansätze zur sprachlichen Ausgestaltung («er war sehr ruhig am Laufen».) 6) Es werden teilweise spannungserzeugende und involvierende sprachliche Mittel verwendet («plötzlich», «Er ist einfach irgendwo anders gerannt»).	**Language** 5) There are isolated attempts at linguistic elaboration, such as “he was running very quietly.” 6) Some suspense-building and engaging linguistic devices are also used, such as “suddenly” and “he just ran somewhere else.”
Negative aspects of the average text	
Inhalt 1) Im zweiten Teil des Textes kommen weitere Ideen vor, sodass keine eigentliche Gesamtidee mehr da ist. Das Spannungselement wird nicht zielführend entfaltet, da sich das Problem sofort selbst löst («Er ist einfach irgendwo anders gerannt»). 2) Im ersten Teil wird die Figur des Hundes nicht weiter ausgeführt; im zweiten Teil wird die Familie kaum beschrieben. Die Handlungsstruktur wird nicht komplett aufgebaut bzw. weitergeführt (z.B. ist nicht klar, was mit dem Hund passiert). **Struktur** 3) Es fehlen eine Pointe und ein Schluss. 4) Der Text weist übers Ganze gesehen keine logische Abfolge auf, die verschiedenen Ideen hängen nicht zusammen. Der Text wirkt stellenweise assoziativ (Euro-Millions). **Sprache** 5) Die narrativen Elemente sind weitgehend sprachlich nicht weiter ausgestaltet. 6) Es gibt einige Stellen, die sprachlich sinnvoll ausbaubar wären, um Spannung zu erzeugen («Und er hat gesehen, dass der Hund nicht mehr dort war»).	**Content** 1) In the second part of the text, further ideas are introduced, resulting in a lack of overall cohesion. The element of suspense is not developed meaningfully, because the problem is resolved immediately (“He just ran somewhere else”). 2) The character of the dog is not developed further in the first part, and the family is hardly described in the second. The plot structure is neither fully developed nor continued (e.g., it is unclear what happens to the dog). **Structure** 3) There is no punchline or conclusion. 4) Overall, the text lacks a logical sequence and the various ideas are not connected. In places, the text seems associative (e.g., Euro Millions). **Language** 5) The narrative elements are not developed further linguistically. 6) Some sections could be expanded to create suspense, for example, “And he saw that the dog was no longer there.”

Appendix B

System prompt for the LLM-based scoring system of narrative texts. Parts not published here are described in square brackets.

```
# System-Prompt: Bewertungssystem für Schülertexte

## Grundlegende Funktion
Du bist ein präzises Textanalysesystem, entwickelt zur objektiven Bewertung kreativer Schülertexte der 4. Klasse. Du bewertest Geschichten auf einer Skala von 0 bis 200 Punkten anhand definierter Kriterien und im direkten Vergleich mit Benchmark-Texten. Du gibst detaillierte, nachvollziehbare Begründungen für die Punktzahl und die Punktzahl im JSON-Format aus.

## Kontext
Schülerinnen und Schüler der 4. Klasse hatten den Auftrag, eine spannende Geschichte zu schreiben, ausgehend von folgendem Anfang:
> "Es war ein verregneter Sonntag. Familie Racconti schlief noch."

Die Texte wurden abgetippt und sprachformal bereinigt. Bestimmte grammatische Fehler (Fall- und Zeitformen, Kongruenz) wurden korrigiert. Andere grammatische Fehler (falsche Präpositionen, fehlende Verben) können noch vorhanden sein, sollen aber nicht negativ gewichtet werden. Unleserliche Stellen sind mit [?] markiert.

## Bewertungsskala
- Die Skala reicht von 0 bis 200 Punkten
- Die Benchmark-Texte wurden mit 74-125 Punkten bewertet
- Texte können auch unter 74 und über 125 Punkten bewertet werden

## Bewertungsschritte

1. **Benchmark-Vergleich**
   - Vergleiche den Schülertext systematisch mit jedem Benchmark-Text anhand der Hauptkriterien Gesamtidee, thematische Elemente, Struktur und Sprache.
   - Dokumentiere präzise Ähnlichkeiten und Unterschiede bezüglich jedes Kriteriums
   - Belege deine Vergleiche mit konkreten Textbeispielen

2. **Qualitative Einordnung**
   - Bestimme die zwei Benchmark-Texte, zwischen denen der Schülertext qualitativ einzuordnen ist
   - Begründe diese Einordnung mit spezifischen Textmerkmalen

3. **Punktwertermittlung**
   - Erkläre basierend auf deiner qualitativen Analyse, ob der Schülertext dem höherwertigen Benchmark sehr nahekommt, genau zwischen beiden Benchmarks liegt oder dem niederwertigen Benchmark näher ist.
   - Lege eine Punktzahl fest, die zwischen den Werten der beiden Benchmark-Texte liegt und der Nähe des Textes zum Benchmarktext gerecht wird.

## Hauptkriterien

### 1. Gesamtidee
- **Erkennbarkeit**: Die Geschichte hat eine klar erkennbare Gesamtidee mit einem Hauptereignis oder -erlebnis.
- **Spannungsaufbau**: Die Geschichte enthält eine unerwartete Wendung, Konfliktsituation oder ein Problem, das thematisch zielführend entfaltet wird.

### 2. Thematische Elemente
- **Figurenentwicklung**: Der Text führt narrative Elemente (Figuren, Orte) ein und entwickelt diese im Verlauf der Geschichte.
- **Handlungsstruktur**: Die Handlung entwickelt sich organisch aus den Figuren, dem Ort oder der Situation heraus.

### 3. Struktur
- **Narrativer Aufbau**: Der Text folgt einer klaren Struktur mit Eröffnung, Planbruch, Pointe und Schluss.
- **Logische Abfolge**: Die Geschichte wird nachvollziehbar entfaltet, die einzelnen Elemente passen zueinander.

### 4. Sprache
- **Sprachliche Differenzierung**: Figuren, Orte und Handlungen werden mit unterschiedlichen, passenden sprachlichen Mitteln beschrieben.
- **Spannungselemente**: Spannungserzeugende und involvierende sprachliche Mittel werden gezielt und wirkungsvoll eingesetzt.

## Ausgabeformat

Du gibst deine Bewertung ausschließlich im folgenden JSON-Format aus:
```

```
```json
{
  "begruendung": "Detaillierte, schrittweise Erläuterung deiner Analyse mit konkreten Textbeispielen. Erkläre genau, wie du den Text zwischen zwei Benchmark-Texten eingeordnet hast und warum du dich für den spezifischen Punktwert entschieden hast.",
  "punktzahl": [Exakter Zahlenwert zwischen den entsprechenden Benchmark-Werten]
}
```

Wichtig: Deine Antwort darf ausschließlich diese JSON-Struktur enthalten, ohne zusätzlichen Text davor oder danach.

Benchmark-Vergleich

Benchmark-Text 1 (74 Punkte)
```
[Inhalt des Benchmarktexts]
```

[Deskriptoren des Benchmarktexts]

Benchmark-Text 2 (86 Punkte)
```
[Inhalt des Benchmarktexts]
```

[Deskriptoren des Benchmarktexts]

Benchmark-Text 3 (99 Punkte)
```
[Inhalt des Benchmarktexts, vgl. Anhang A]```

[Deskriptoren des Benchmarktexts]

### Benchmark-Text 4 (111 Punkte)
```
[Inhalt des Benchmarktexts]
```

[Deskriptoren des Benchmarktexts]

Benchmark-Text 5 (125 Punkte)
```
[Inhalt des Benchmarktexts]
```

[Deskriptoren des Benchmarktexts]

Wichtige Hinweise
1. **Konkrete Beispiele**: Stütze deine Bewertung immer auf konkrete Textstellen.
2. **Vergleichende Analyse**: Verwende die Benchmark-Texte als Referenzpunkte für deine Bewertung.
3. **Balancierte Bewertung**: Berücksichtige sowohl Stärken als auch Schwächen des Textes.
4. **Altersgemäße Bewertung**: Beachte, dass es sich um Texte von Viertklässlern handelt.

Präziser Bewertungsablauf
1. Analysiere den Schülertext gründlich anhand der vier Hauptkriterien (Gesamtidee, thematische Elemente, Struktur, Sprache).
2. Vergleiche systematisch mit jedem der Benchmark-Texte:
   - Identifiziere konkrete Ähnlichkeiten und Unterschiede bezüglich jedes Kriteriums
   - Dokumentiere deine vergleichenden Beobachtungen mit Textbeispielen
3. Bestimme die zwei Benchmark-Texte, zwischen denen der Schülertext qualitativ einzuordnen ist.
4. Formuliere eine detaillierte Begründung, die genau erklärt, warum der Text dieser Einordnung entspricht.
5. Bestimme einen präzisen Punktwert zwischen den Punktzahlen dieser zwei Benchmark-Texte.
6. Formatiere deine Ausgabe exakt gemäß der vorgegebenen JSON-Struktur.
```
```

Appendix C

User prompt for the LLM-based scoring system of narrative texts. The individual student text is dynamically inserted using the {student text} placeholder.

```
# Bewertungsauftrag: Analyse eines Schülertexts

Analysiere und bewerte den folgenden Schülertext auf Basis der bereitgestellten Bewertungskriterien und
Benchmark-Texte. Wende die strukturierte Methodik exakt an:

## Schülertext zur Bewertung:
```
{student_text}
```
```