HFT at BEA 2026 Shared Task 2: Blunt-Edge Models for Hybrid Grading

Ulrike Padó
Hochschule für Technik (HFT) Stuttgart
Schellingstr. 24
70174 Stuttgart, Germany
ulrike.pado@hft-stuttgart.de

Abstract

Open-source LLMs with simple, zero-shot prompts are at best middling graders on the BEA 2026 Automated Grading Shared Task – blunt-edge models, in fact. However, they are good enough to support human graders and save them time. We demonstrate the application of a hybrid grading approach that first transparently defines the success criteria and then pairs a zero-shot LLM grader with human review. The hybrid approach outperforms the LLM grader on its own and has the added advantage of keeping the human in the loop.

1 Introduction

Short answer questions are a welcome opportunity for educators to probe their students' ability to explain or analyze a problem. By phrasing an answer in their own words, the students prove their factual knowledge, show their thought processes and also demonstrate their language ability in language instruction settings. However, manual grading of short answer questions, while yielding useful insights into common misconceptions or errors, is time-consuming. Therefore, many attempts at automating short answer grading have been made over the years, and have profited strongly from the availability of common benchmark data.

The BEA 2026 Short-Answer Grading Shared Task (Gombert et al., 2026) offers a new data set for German rubric-based scoring. We approach this data with LLM grading models that deal gracefully with rubric instructions. We specifically choose open-source models to ensure independence of commercial offerings: These are prone to hidden changes or complete retraction and therefore endanger mid- to long term replicability and pose risks to data privacy in practical use.

Also, we intentionally limit ourselves to the "blunt edge" of research: Instead of creating an optimized "cutting-edge" model, we stress low cost and broad re-usability for other tasks by evaluating a simple zero-shot approach using open-source LLMs and standardized prompts from the literature. Our blunt-edge model performs close to more sophisticated offerings in the information-poor Unseen Question setup, but clearly lags behind approaches that integrate the available information in the Unseen Answer setting (see Section 3.3).

To address imperfect model performance, we propose using a hybrid human-machine grading process that integrates human decision making with automated grading in such a way that the most reliable machine grade predictions are retained and the less reliable ones revised manually. The performance requirements of the process are defined by the educator beforehand (Padó et al., 2024). The simulated output of the hybrid process is competitive with cutting-edge models, especially when no sample answers are available (Section 4).

Finally, we argue that partial (or full) human review of the proposed model grades not only allows the hybrid process to reach the pre-defined performance requirements; it also gives the educator insights into the student data. In fact, human oversight over grading decisions is crucial for high-stakes systems in practical use in the European Union: The recent EU AI Act requires extensive risk management and documentation of any AI system that grades student submissions without human intervention.[1] In contrast, this explicitly does not apply to systems that just prepare data for a human evaluation process. [2]

We now describe our approach in more detail, starting with the model and prompt selection for our blunt-edge model on the Shared Task trial data, its performance in the competition proper and the expected outcome of the hybrid process on the (unseen) Shared Task training data.

[1] Such systems are classed as high-risk systems in the AI Act, Chapter 3, Section 1, Article 6(3) and Annex III, (3)

[2] Cf. AI Act, Chapter 3, Section 1, Article 6(3d)

2 Related Work

Short-Answer Grading is a well-established task (see the overviews of Burrows et al., 2015; Bai and Stede, 2023). While most of the research is on English data, several German data sets have been collected over the last 15 years, e.g., CREG (Meurers et al., 2011), CSSAG (Padó and Kiefer, 2015), ASAP-DE (Horbach et al., 2018) and SAF-DE (Filighera et al., 2022). However, especially the older data sets are limited in size, making the BEA 2026 Shared Task 2 data a welcome addition.

Along with every subfield of Natural Language Processing, Short-Answer Grading for German has seen a surge of experimentation with LLMs. Filighera et al. (2022) used a Transformer model with a trained classification head on the SAF-DE data and saw up to 85% prediction Accuracy in cases where the question had been seen during training (but the test answers had not) and 55% for unseen questions. With a similar approach, Padó (2016) achieved 84.4% prediction Accuracy on the CSSAG data on seen questions and 69% on unseen questions. These numbers are similar to non-neural benchmarking results for CREG (85% seen-question Accuracy, Meurers et al. 2011).

Using one-shot prompting that provided the LLM with the correct reference answer and the expected range of points, Speiser and Weng (2024) were able to vastly improve Accuracy on unseen questions from CSSAG to 87%. Note that Metzler et al. (2024), working on English data and a small German student answer data set in the statistics domain, recommend prompting using rubrics over n-shot answer examples, because they find that while n-shot prompting improve results, it also leads to inconsistent predictions.

In sum, prompting LLMs is a promising approach to solving the automatic grading task, especially since they can make use of the information in rubrics without further pre-processing, and appear to show more consistent performance with rubrics. However, the literature also shows that providing as much information as possible in the form of reference answers, sample answers or the expected point distribution improves prediction performance.

Regarding the human-machine hybrid grading strategy, a number of proposals have been made over time to integrate machine predictions with human grading or revision, in order to lighten the human workload. Some approaches cluster student answers to reduce the need for labeling to one instance per cluster (Basu et al., 2013; Zehner et al., 2016), with the advantage of allowing teachers insight into their students' frequent error patterns and potentially giving feedback per cluster.

Human review of some part of the machine predictions without clustering, based on the confidence of the grading models and their strengths and weaknesses, is also proposed by Schneider et al. (2023) or Kwako et al. (2026). Alternatively, Speiser and Weng (2024), like Vittorini et al. (2021), advocate for human-machine dual grading with revisions only in case of disagreement, which sacrifices workload reduction for full human review and improved correctness. However, Vittorini et al. (2021) still find a 40% speedup for reviewing over grading from scratch.

An investigation of very restricted, open-source LLMs to solve a Shared Task was undertaken by Góngora et al. (2025) at the BEA Shared Task 2025 (Pedagogical Ability Assessment of AI-powered Tutors): They only looked at models with 1b parameters or less to demonstrate the limits of what is possible with restricted access to computation resources. Padó (2026) approaches the 2025 Shared Task with the same blunt-edge, hybrid grading process methodology applied here, and finds that the hybrid process can be successfully applied to reach a pre-defined Accuracy threshold.

3 Shared Task: Grading Models

3.1 Method

We experiment with the open-source LLMs available to us for research purposes through Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen. We work with the seven text-based models that responded in less than 30s per API request and consistently produced the standardized JSON output format needed for post-processing. These were text-based models `gpt-oss`[3], `qwen3-32b`[4], `gemma-3-27-it`[5], `mistral-large-3-675b-instruct-2512`[6] and `llama-3.3-70b-instruct`[7], specialized reasoning model `qwen3-30b-a3b-thinking-2507`[8]

[3] https://huggingface.co/openai/gpt-oss-120b
[4] https://huggingface.co/Qwen/Qwen3-32B
[5] https://huggingface.co/google/gemma-3-27b-it
[6] https://huggingface.co/mistralai/Mistral-Large-3-675B-Instruct-2512
[7] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[8] https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

and `llama-3.1-sauerkrautlm-70b-instruct`[9], which was fine-tuned on German-English data. All models were run at a temperature of 0.01, so in conservative mode. We used the RAGAS[10] libraries for implementation.

The second parameter in our setup is the nature of the prompt. We use four prompts in a zero-shot setting (i.e., the model receives instructions, but no examples), cf. Padó (2026). Our *Baseline* prompt just instructs the LLM to return "Correct" if the student answer is correct given the rubric, and "Incorrect" otherwise. Next, a prompt proposed by Fan et al. (2025) in the BEA 2025 tutor evaluation task goes one step further and explains the task context (a student answering a question and receiving a grade). An alternative prompt specifies a *Role* for the LLM (it simulates an experienced STEM teacher). Finally, we also use a prompt template by Wang et al. (2024), which instructs the model to simulate a committee of experts solving the task (the prompt was shortened to one sample conversation). All the prompts were in German (translated from the originals as needed) and can be found in the Appendix. With this set of progressively more complex prompts, we want to probe how much information the LLMs need to grade well.

3.2 Model and Prompt Selection

We evaluate our seven models and four prompts on the Shared Task trial data in the two-way setting (labels "Correct" and "Incorrect"). Our zero-shot approach mimics the Unseen Question task. Table 1 shows the results. As expected, the sparse Baseline **prompt** rarely yields best results, but neither does the very complex Wang et al. prompt (Mistral, SauerkrautLM and Gemma even do much worse than for the other prompts). Instead, models do best with the shorter Fan et el. and Role prompts that provide some context to the task but are essentially simple. Variation in our LLMs' predictions, accounts for as much as 0.025 points QWK difference across five runs for `LLama+Fan et al.` (max. 0.49, min. 0.46, average 0.48) and 0.016 points QWK across five runs for `GPT-OSS+Fan et al.` (max. 0.48, min. 0.46, average 0.48). Given this observation, performance is surprisingly consistent across the Baseline, Fan et al. and Role prompts and across different models. This is positive because it shows a certain robustness to model identity and details of prompting that promises that performance will carry over to other configurations.

In an unplanned experiment, a correction of the training and trial data for the 2-way task gave insight into the importance of covering the whole answer space with rubrics. When the initial data set that had only the rubrics for "Incorrect" and "Correct" was replaced with a revised data set that also contained the "Partially Correct" rubrics (to count as "Incorrect"), grading performance went up by around 0.03 points QWK across all models and prompts. This was carried by improved Precision for the "Correct" predictions at almost unchanged Precision for label "Incorrect"; clearly, some of the "Partially Correct" answers had initially been erroneously identified as "Correct" due to the partial overlap with the "Correct" rubrics.

3.3 Results on Shared Task Test Set

For evaluation on the test data, we chose the best-performing models and submitted several runs each in order to account for prediction variation. The best-performing and most consistent models are GPT-OSS and Llama. We paired both models with the Fan et al. prompt, the simplest successful prompt.

We submitted model predictions for the 2-way Unseen Question and Unseen Answer tasks. Since we use a zero-shot setting, we expect to see the same performance on both test tasks as on the trial data reported above - around 0.48 QWK.

Table 2 shows that this expectation was borne out by the `GPT-OSS+Fan et al.` model. `Llama+Fan et al.` did somewhat worse, although performance on the trial data had generally been similar. Still, the blunt-edge models again show remarkable robustness, this time across data sets, and achieve 5th place (out of 9) in the 2-way UQ task.

The competition results on the test data show that the largest performance gains of the cutting-edge models are in the Unseen Answer setting, where task-specific information is available (Gombert et al., 2026). Here, the blunt-edge model places last. Even though the rubrics provide it with some task-specific information, more knowledge about the task in the shape of previously graded answers remains hugely beneficial – a familiar pattern in ASAG independent of grading algorithm. We see this situation as a trade-off: If data, development time and computing resources are available and models can be re-used often, the development of a cutting-edge approach is clearly worth it. However,

[9] https://huggingface.co/VAGOsolutions/Llama-3.1-SauerkrautLM-70b-Instruct

[10] https://docs.ragas.io/en/latest/

Model	Baseline	Fan et al.	Role	Wang et al.
GPT-OSS	0.4772	0.4833	0.4759	**0.4846**
Llama	0.4772	**0.4927**	0.4846	0.4538
Mistral	0.4632	0.4366	**0.4632**	0.3428
Qwen3-32b	0.4621	**0.4840**	0.4832	0.4207
Qwen3-thinking	0.4535	0.4335	**0.4602**	0.4595
SauerkrautLM	**0.4464**	0.4453	0.4171	0.3290
Gemma	0.4056	0.4181	**0.4351**	0.3784

Table 1: QWK for LLMs and four prompts on the trial data, two-way Unseen Question task (best result in bold).

	2-way UQ			2-way UA		
Model	QWK	F_1	Pos.	QWK	F_1	Pos.
IWM-DKM	0.55	0.815	1	0.726	0.887	1
GPT-OSS	0.482	0.788	5	0.477	0.783	9
Llama	0.452	0.767	–	0.435	0.766	–

Table 2: QWK and weighted F_1 for the two-way Unseen Question (UQ) and Unseen Answer (UA) test data. Task winner and our models (Fan et al. prompt), showing the best submitted result and the team rank (out of nine).

in many teaching settings questions change from exam iteration to exam iteration with no sample answers available; in these cases, the blunt-edge models are a viable, low-cost alternative.

4 Hybrid Grading

As expected, the zero-shot open-source models on their own do not perform competitively on the Shared Task test data, especially in the information-rich Unseen Answer setting. However, we argue that they are fully sufficient to support hybrid human-machine grading when following our four-step grading process (cf. Padó et al. 2024). The process guides an educator through model selection and hybrid grading, by identifying the concrete grading model's "preferred" class, which it predicts best. Accepting predictions for this class and reviewing all predictions of the other class(es) reduces human workload and significantly improves grading results over machine grading alone. This model-based selection of presumably reliable predictions precludes the need for confidence scores for each grade, which may not be available.

We now step through the process and apply it to the (unseen) Shared Task training data.

Step 1: Define Quality Threshold Since the maximum acceptable error for the output of the grading process can vary for each usage context, the hybrid grading process calls for defining it explicitly first. Published short-answer corpora show 10-15% disagreement (lower for higher stakes) among human graders (Mieskes and Padó, 2018; Nazaretsky et al., 2022). Therefore, we accept 15% error as the expected human performance.

Step 2: Choose Data Set The next task is to identify data from one's own setting with existing human grades to test the grading models on. On the basis of this evaluation, the final grading setting for the hybrid grading process is chosen. This step recognizes that model performance can vary widely between data sets and that benchmark performance cannot be expected to carry over to different settings. We use the Shared Task trial data.

Step 3: Choose Grading Model We use our results for the trial data from 3.2. At this step in the process, however, we drill further down into the two best models' performance to identify their "preferred" classes and class-specific performance. We look for high Precision on at least one class – optimally, at least the accepted error threshold for the task (cf. Step 1) to ensure acceptable error for each class. Our two best models are close to this threshold: `GPT-OSS+Fan et al.` shows $Prec_{Incorrect} = 83.00$ and `Llama 3.3+Fan et al.` has $Prec_{Incorrect} = 84.3$.

Step 4: Decide on Use Given the results in Step 3, we decide to use `Llama 3.3+Fan et al.` to grade the Shared Task training data, which are unseen to our zero-shot model. We will accept any "Incorrect" grades and assume that human review fixes all errors in the "Correct" grade predictions.

On the 7072 answers in the training set, the model made 5103 predictions of class "Incorrect" (to be accepted as-is) and 1969 predictions of class "Correct" (to be revised). This means a reduction in human workload by 71%.

The gold labels show that 840 predictions of "Incorrect" were in fact wrong. If the revised labels are assumed to be perfect, this makes for a remaining error of 12%, well within our margins. Weighted F_1 for the hybrid process would be 0.896 in this best-case scenario, which is higher than the best reported F_1 score in the competition.

If we (more realistically) assume that the human expert makes evenly-distributed errors for 15% of the 1969 labels during revision, missing some true errors and marking some truly correct answers as incorrect, overall error for the hybrid process is 16% (still close to our threshold) and weighted F_1 drops to 0.828 – eighth place in the Unseen Answers track, and first in the information-poor Unseen Questions track. This makes the hybrid process a clear recommendation for cases where little is known in advance about the questions and answers: In such cases, the simple blunt-edge model saves development time and resources, while the hybrid setup ensures competitive performance and allows the human grader insights into students' answer patterns.

5 Discussion and Conclusions

We have presented our contribution the the BEA 2026 Automated Grading Shared Task: A blunt-edge, but replicable and relatively low-cost LLM grader that relies on off-the-shelf open source LLMs and simple zero-shot prompts.

While the grader holds its own in the information-poor Unseen Questions task, it clearly lags behind the cutting-edge models in the Unseen Answer task. However, we argue that it is good enough to be combined with human judgments in a hybrid grading process which accepts the model grades for the class predicted with higher Precision and presents the other grades to human review. This hybrid process even outperforms the cutting-edge models in simulations on unseen data, especially in information-poor settings.

Note that the hybrid strategy is applicable to any grading model, and further performance gains are possible with better machine graders. For any grading task that has little previously-annotated training data, we still explicitly recommend the use of blunt-edge models, however. The use of off-the-shelf open source LLMs ensures data privacy, replicability and reduction of the tasks' energy footprint, and blunt-edge performance is robust in information-poor settings.

In any case, the manual revision step allows the human graders insight into answer patterns that are not available in fully automated grading.

Importantly, the hybrid grading process has a bias towards whichever label is predicted with less error by the grading model and therefore accepted without review. This is often the majority class in the training data, especially for trained or fine-tuned models. For the Shared Task data, this class is "Incorrect". This means that all the remaining error after application of the hybrid process is concentrated in cases where the machine erroneously predicts a student to be wrong. In a summative setting, this is a worst-case scenario. In a formative setting, it can be acceptable depending on the amount of feedback a student receives and the opportunities they have to clear up confusion caused by inaccurate machine grades.

Note, however, that depending on the data set and grading model, the process bias can well be inverted (towards "Correct"). This is why we strongly advocate for experimenting with the grading model on realistic data before choosing the manual-machine balance for the task in question. For example, educators could choose complete manual review for high-stakes tests or in case of poor model performance on their data, or on the opposite end of the spectrum skip manual review completely if grading a low-stakes feedback activity and using a reliable model with an acceptance bias for "Correct". In fact, the hybrid process has the advantage of allowing this human-machine balancing on an individual basis while taking into account the testing situation, the expected reliability of the machine grades, the available human work time and individual preferences for manual oversight.

Naturally, to fully avoid unfair bias in all situations, all machine grades have to be revised every time. This can still be worthwhile, since Vittorini et al. (2021) find a time saving of 40% for revision over grading from scratch. Complete revision also ensures that the requirements of the EU AI Act for the use of AI in grading are met. For both these reasons, full review is recommended in high-stakes situations.

Limitations

Our approach has been simulated for different data sets, but a full trial with real human annotators is is still outstanding. To our understanding, the partial review process does not satisfy the requirements for low-risk AI use according to the EU AI Act. For this reason, additional experiments on the time savings of full vs. partial review should also be done. No additional inquiries into the effect of adding more task information (for example through few-shot prompting) in the Seen Question task were made.

Ethics

Our approach minimizes, but does not eliminate, the energy footprint of LLM usage. There is a possible bias in the resulting grades since errors concentrate in one predicted class. Also, there are currently no checks for other machine grading biases, e.g., against students who are not native speakers, cf. Padó et al. (2024)).

Acknowledgments

The author gratefully acknowledges the LLM hosting services granted by the KISSKI project of Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG). Many thanks also to two anonymous reviewers for their constructive comments.

References

Xiaoyu Bai and Manfred Stede. 2023. A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *Int. J. Artif. Intell. Educ.*, 33(4):992–1030.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25:60–117.

Yuming Fan, Chuangchuang Tan, and Wenyu Song. 2025. BJTU at BEA 2025 shared task: Task-aware prompt tuning and data augmentation for evaluating AI math tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1073–1077, Vienna, Austria. Association for Computational Linguistics.

Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. 2022. Your answer is incorrect... would you like to know why? Introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8577–8591.

Sebastian Gombert, Zhifan Sun, Fabian Zehner, Jannik Lossjew, Tobias Wyrwich, Berrit Katharina Czinczel, David Bednorz, Sascha Bernholt, Knut Neumann, Ute Harms, Aiso Heinze, and Hendrik Drachsler. 2026. Report on the BEA 2026 Shared Task on Rubric-based Short Answer Scoring for German. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

Santiago Góngora, Ignacio Sastre, Santiago Robaina, Ignacio Remersaro, Luis Chiruzzo, and Aiala Rosá. 2025. RETUYT-INCO at BEA 2025 shared task: How far can lightweight models go in AI-powered tutor evaluation? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1135–1144, Vienna, Austria. Association for Computational Linguistics.

Andrea Horbach, Sebastian Stennmanns, and Torsten Zesch. 2018. Cross-lingual content scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 410–419.

Alexander Kwako, Susan Lottridge, and Christopher Ormerod. 2026. Using confidence modeling to optimize overall score quality in hybrid scoring systems. *Educational Measurement: Issues and Practice*, 45(2):e70019.

Tim Metzler, Paul G. Plöger, and Jörn Hees. 2024. Computer-assisted short answer grading using large language models and rubrics. In *INFORMATIK*, Lecture Notes in Informatics (LNI). Gesellschaft für Informatik.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(4):355–369.

Margot Mieskes and Ulrike Padó. 2018. Work smart - reducing effort in short-answer grading. In *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning*, pages 57–68.

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):914–931.

Ulrike Padó. 2016. Get semantic with me! The usefulness of different feature types for short-answer grading. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2186–2195.

Ulrike Padó. 2026. Hybrid evaluation of tutor dialogues. In *Agentic learning Ecoystems and ITS, 22nd International Conference (ITS 2026)*. Springer Nature Switzerland. To appear.

Ulrike Padó, Yunus Eryilmaz, and Larissa Kirschner. 2024. Short-answer grading for German: Addressing the challenges. *International Journal of Artificial Intelligence in Education*, 34(4).

Ulrike Padó and Cornelia Kiefer. 2015. Short answer grading: When sorting helps and when it doesn't. In *Proceedings of the Workshop on NLP for Computer-Aided Language Learning*, pages 42—50, Vilnius, Lithuania.

J. Schneider, R. Richner, and M. Riser. 2023. Towards trustworthy autograding of short, multi-lingual, multi-type answers. *Int J Artif Intell Educ 33*, 33:88–118.

Sebastian Speiser and Annegret Weng. 2024. Enhancing short answer grading with OpenAI APIs. In *IEEE International Conference on IT in Higher Education and Training (ITHET)*.

Pierpaolo Vittorini, Stefano Menini, and Sara Tonelli. 2021. An AI-based system for formative and summative assessment in Data Science courses. *International Journal of Artificial Intelligence in Education*, 31:159–185.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Fabian Zehner, Christine Sälzer, and Frank Goldhammer. 2016. Automatic coding of short text responses via clustering in educational assessment. *Educational and psychological measurement*, 2(76):280–303.

A Appendix: Prompts

Baseline "Bewerte mit 1, wenn die Antwort gegeben die Bewertungsvorschrift richtig ist. Sonst bewerte mit 0."

Fan et al. "Ein Schüler beantwortet eine Frage. Deine Aufgabe ist es, zu bewerten, ob die Antwort auf die Frage richtig ist. Die Bewertungsvorgaben für Richtig und Falsch liegen dir vor. Deine Aufgabe ist, zu bewerten, ob die Antwort gegeben die Bewertungsvorschriften richtig ist. Bewerte mit 1, wenn die Antwort richtig ist. Sonst bewerte sie mit 0."

Role "Du bist ein erfahrener Lehrer in den Naturwissenschaften. Du bewertest eine Schülerantwort auf eine Frage. Du hast Bewertungsrichtlinien gegeben. Bewerte die Antwort mit 1, wenn sie den Richtlinien entsprechen. Sonst bewerte sie mit 0."

Wang et al. "Wenn du eine Aufgabe bekommst, beginne damit, die Beteiligten zu identifizieren, die zur Lösung beitragen werden. Dann starte einen mehrstufigen Zusammenarbeitsprozess, bis eine endgültige Lösung gefunden ist. Die Beteiligten geben dir kritische Rückmeldungen und detaillierte Vorschläge, wenn nötig. Hier sind einige Beispiele:

Beispiel-Aufgabe 1: Verwende Zahlen und grundlegende Rechenoperationen (+ – * /), um 24 zu erhalten. Du musst alle Zahlen verwenden, und jede Zahl darf nur einmal genutzt werden. Eingabe: 6 12 1 1 Beteiligte: Du (KI-Assistent); Mathematik-Experte Start der Zusammenarbeit! Mathematik-Experte: Lass uns die Aufgabe genau analysieren. Du musst sicherstellen, dass du die Anforderung erfüllst: Du musst genau die vier Zahlen (6 12 1 1) verwenden, um 24 zu bilden. Um 24 zu erreichen, kannst du an die häufigen Teiler von 24 denken, wie 4, 6, 8 oder 3, und versuchen, diese zunächst zu konstruieren. Außerdem solltest du potenzielle Summen in Betracht ziehen, die 24 ergeben, wie zum Beispiel 12 + 12. Du (KI-Assistent): Danke für die Hinweise! Hier ist eine erste Lösung: (12 / (1 + 1)) * 6 = 24 Mathematik-Experte: Lass uns die Antwort Schritt für Schritt überprüfen. (1 + 1) = 2, (12 / 2) = 6, 6 * 6 = 36 – das ist nicht 24! Die Antwort ist falsch. Kannst du das durch andere Kombinationen korrigieren? Bitte vermeide ähnliche Fehler. Du (KI-Assistent): Danke, dass du den Fehler aufgezeigt hast. Hier ist eine überarbeitete Lösung, bei der berücksichtigt wird, dass 24 auch durch 3 × 8 erreicht werden kann: (6 + 1 + 1) × (12 / 4) = 24. Mathematik-Experte: Lass uns prüfen, ob die Berechnung stimmt. (6 + 1 + 1) = 8, 12 / 4 = 3, 8 × 3 = 24. Die Rechnung ist korrekt, aber du hast die Zahlen 6, 1, 1, 12 und 4 verwendet – das entspricht nicht der Eingabe 6, 12, 1, 1. Kannst du vermeiden, eine Zahl zu verwenden, die nicht in der Eingabe enthalten ist? Du (KI-Assistent): Du hast recht. Hier ist eine überarbeitete Lösung, bei der berücksichtigt wird, dass 24 auch durch 12 + 12 erreicht werden kann, und ohne zusätzliche Zahlen: 6 × (1 – 1) + 12 = 24. Mathematik-Experte: Lass uns die Antwort nochmal überprüfen. 1 – 1 = 0, 6 × 0 = 0, 0 + 12

= 12. Du bist ganz nah dran. Hier ein Tipp: Versuche, „1 – 1“ durch „1 + 1“ zu ersetzen. Du (KI-Assistent): Klar, hier ist die korrigierte Antwort: 6 × (1 + 1) + 12 = 24 Mathematik-Experte: Lass uns die Lösung überprüfen. 1 + 1 = 2, 6 × 2 = 12, 12 + 12 = 24. Du hast die Zahlen 1, 1, 6, 12 verwendet – das entspricht genau der Eingabe 6, 12, 1, 1. Alles passt! Zusammenarbeit beendet! Endgültige Antwort: 6 × (1 + 1) + 12 = 24 Identifiziere nun die Beteiligten und lass sie die folgende Aufgabe schrittweise gemeinsam lösen.

Ein Schüler beantwortet eine Frage. Die Bewertungsrichtlinien für Richtig und Falsch liegen dir vor. Bewerte die Antwort mit 1, wenn sie den Richtlinien entsprechen. Sonst bewerte sie mit 0."