IWM-DKM at BEA 2026 Shared Task 2: Supplementing Supervised Fine-Tuning for Rubric-Based Short Answer Scoring

Kate Rebecca Belcher[1,2], **Marius De Kuthy Meurers**[3],
Kordula De Kuthy[1], **Detmar Meurers**[1,2]

k.belcher@iwm-tuebingen.de, marius.dkm@tum.de, k.dekuthy@iwm-tuebingen.de, d.meurers@iwm-tuebingen.de

[1]Leibniz-Institut für Wissensmedien, [2]University of Tübingen, [3]Technical University of Munich

Abstract

In this paper, we present the IWM-DKM team submissions to the BEA 2026 Shared Task 2: Rubric-based Short Answer Scoring for German. We systematically explored how fine-tuned language models can be reliably employed for short answer scoring, for which three aspects turn out to be particularly beneficial: supplementing the fine-tuning process with generated domain expertise, restructured rubrics, and thinking traces. To increase the robustness of the scoring, we combine distinct approaches in an ensemble. Our best submissions finished in first place across all tracks, indicating promise for the further application of these strategies in automatic scoring.

1 Introduction

In contrast to previous challenges for automatic short answer scoring, such as the 2012 Kaggle challenge (Barbara et al., 2012) and the 2013 SemEval Task 7 (Dzikovska et al., 2013), the BEA 2026 Shared Task 2 (Gombert et al., 2026) represented the first challenge in which *scoring rubrics* formed an essential element of the scoring, and the first large-scale non-English short-answer challenge. The shared task offered two labeling schemes: a 3-way classification of answers into 'Correct', 'Partially correct' and 'Incorrect', and a binary 'Correct' / 'Incorrect' classification. Additionally, two evaluation setups were provided: *Unseen Answers* (a held-out set of answers to the questions from the training data), and *Unseen Questions* (answers for new questions not included in the training data). Participating as team IWM-DKM, we submitted entries to all four tracks of the challenge.

We utilized LoRA-fine-tuned LLMs (Hu et al., 2022) from the Qwen family (Yang et al., 2025) as the backbone of our grading system and successfully improved performance by a combination of *augmentation* of the model input with additional relevant information such as *domain-specific background knowledge*, and *restructuring information*, including both the rubrics and the model's thinking traces to increase grading consistency.

Our approach produced the best submission in both *unseen question* (UQ) tracks with a single fine-tuned model combined with several supplementary strategies, and achieved first place in both *unseen answer* (UA) tracks by combining the strengths of multiple models into a majority-vote ensemble. In the following, we discuss related work on rubric-based short answer scoring and what motivated our use of the techniques we implemented. We then describe our methodology in detail and outline the results for each track.

2 Related Work

Recent years have seen an increase in deep learning approaches (e.g. Bexte et al., 2022; Ormerod, 2022) and LLM-implementations for short answer scoring (e.g. Chamieh et al., 2024; Frohn et al., 2025; Gao et al., 2025). In rubric-based scoring, descriptive scoring rubrics are integrated into the auto-grading methodology, acting as a mechanism to guide model decisions closer in line with principles of human assessment. Wang et al. (2019) integrate "key elements" of scoring rubrics with a word-level attention mechanism in a Bi-LSTM neural grading system. In recent approaches with LLMs, both Zhao et al. (2024) and Jiang and Bosch (2024) find that LLM scoring reliability is improved with the addition of rubrics, and Frohn et al. (2025) observe variable performance dependent on the level of rubric detail. Somewhat contrastingly, Pathak et al. (2025) experiment with prompt granularity, and find that LLM grading may become too harsh when rubrics are evaluated in a point-wise manner. Pathak et al. develop a multi-agent LLM approach, utilizing a majority vote ensemble of models. While they observe benefits of an ensem-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1224–1234

ble over a single model, beyond 3 to 4 LLMs results plateau with the addition of more LLMs. Overall, works to date suggest that rubrics offer a promising way to improve performance in short answer scoring, with the additional benefit of greater grounding in principles of traditional human assessment. However, how best to incorporate rubrics into automatic grading has not yet been comprehensively explored. While many recent approaches integrate rubric information via LLM prompting, the extent to which rubrics can be applied to the fine-tuning of pre-trained LLMs remains unclear. With our shared task submission, we aim to target this gap.

3 Approach

The ALICE-LP dataset consists of questions and answers from four science subject areas with authentic responses from secondary school students in Germany. The training set contained 78 questions with 7899 answers, alongside a trial set of 827 answers. Each item contains the question, the student response, and a rubric describing the criteria for each scoring level. Preliminary experiments with several encoder models appeared to reach a performance ceiling at around 0.75 QWK on the 3-way trial set. Consequently, our primary approach instead utilized supervised fine-tuning of decoder-only language models, specifically Qwen3 (4B/8B) and Qwen3.5 (9B) (Yang et al., 2025), incorporating all three item elements into the model input. We opted to use Qwen models, as relatively small-sized yet high performing open-source LLMs, making them more suitable for fine-tuning and deployment in educational settings than closed-source alternatives (Lin et al., 2026).

We employed Low-Rank Adaptation (LoRA) (Hu et al., 2022) targeting all linear projection layers with a rank $r = 16$ and a scaling factor $\alpha = 16$. This fine-tuning strategy was chosen so as to preserve the underlying model as much as possible whilst also taking into account the nature of the task-dataset. Models were trained using the AdamW optimizer (Loshchilov and Hutter, 2017) with a peak learning rate between 1×10^{-4} and 2×10^{-4} and a global batch size of 16 or 32. We applied a cosine learning rate schedule to smoothly decay the learning rate and trained the models for up to 10 epochs, ultimately selecting the best checkpoint based on the highest validation Quadratic Weighted Kappa (QWK). For certain models, we experiment with different fine-tuning strategies, including Direct Preference Optimization (DPO) (Rafailov et al., 2023) (where the probabilities of correct labels are increased relative to dispreferred options); Fast Gradient Method (FGM) (Goodfellow et al., 2014; Miyato et al., 2017) (an adversarial training method whereby embeddings are perturbed to increase robustness); and NEFTune (Jain et al., 2023) (a further method to introduce noise to the vector embeddings). We indicate in Section 4 for which submissions we employed alternative fine-tuning strategies. We build on this foundation with a range of approaches that aim to specialize the model more closely to the domain of the task and align the model's decision process with the information contained in the rubrics. We do this by providing the model with *more* relevant information and by restructuring the format of it.

3.1 Checklist Thinking

Based on the principle of walking through intermediary steps before giving a final answer, the idea of "thinking" in LLMs has evolved from enhancing prompts with chain-of-thought prompting (Kojima et al., 2022) to integrating thinking steps into the model fine-tuning process (Zelikman et al., 2022), with the aim of improving model consistency. To apply this principle to short answer grading, we introduce a structured reasoning format termed *Checklist Thinking*. Instead of predicting the score label directly, the model is trained to first output a concise set of "Yes/No" checks corresponding to each rubric level within `<think>` tags. We found that a dynamic "Applicable-last" order was most effective—where the model explicitly evaluates and rejects the non-applicable score categories before confirming the criteria for its final predicted label. This method was particularly advantageous in improving recall for the "Partially correct" category, an example of which is:

```
<think>
Incorrect: NEIN
Correct: NEIN
Partially correct: JA
</think>
Partially correct
```

3.2 Rubric Reframing

LLMs typically perform better on tasks that use a structured, closed format. In question answering tasks, a performance drop is observed when moving from multiple choice to open-question tasks (e.g. Li et al., 2024). Several works address this potential challenge for LLMs by breaking down

complex tasks into a series of sub-tasks (Dua et al., 2022; Qin et al., 2024; Zhou et al., 2023), or by decomposing instructions into Yes/No checklists, resulting in greater alignment with human preference (Cook et al., 2024). Taking inspiration from these techniques, we consider how the format of the rubrics can be restructured to be more accessible to LLMs. For each 3-way rubric, we prompt GPT5.1 (Singh et al., 2025) to transform the rubric from high-level descriptions of the categorical levels to a set of questions that can be followed as a decision tree, providing greater structure to guide the LLM to a grading decision. The prompt used to generate the reformulated rubrics is shown in Appendix C.1 and an example is shown in Appendix B.

3.3 Background Knowledge

A characteristic of the dataset is that many questions require an understanding of subject-specific terminology and processes, or relate to external material (e.g., tables, experiments). To provide the necessary domain expertise, and counter some of the lack of context present in the question alone, we supplement the model input with generated "background knowledge". We use Claude Sonnet 4.6 (Anthropic, 2026b) to generate a short textbook-style summary based on the question, the rubric, and all examples of correct student answers available for the given question. The exact number of examples used to generate background knowledge was variable, with an average of 26.46 correct answers per question in the unseen answers train set. For *unseen questions*, where ground-truth labels are absent, we employ a two-step pipeline: We initially generate context from all answers, from which labels are then predicted, and then re-generate background knowledge in a second pass using only the answers labelled as "Correct" (i.e., 'silver-standard' labels) (see Appendix C.2 for prompt details, and Appendix B for an example).

3.4 Skilled-based Categories

To better understand the nature of the data, and due to the breadth of question types in the dataset, we manually annotated each question with the operator used in the question (e.g., *describe*, *justify*, *explain* etc.). In addition, we assign each question to one of four categories according to the skills targeted in the question: *Concept Description* (describing abstract concepts or processes), *Example Led* (analyzing examples in relation to fundamental concepts), *Idea Seeking* (the formulation of hypotheses or examples), and *Mathematical* (calculations and graph interpretations). The conceptual basis for the categories was established in discussion among the authors and annotation was subsequently carried out by one of the authors. An example question for each category can be found in Appendix D. We use these categorizations in two ways. Firstly, motivated by the idea that scoring performance may be variable in accordance with question type (e.g. Meurers et al., 2011), we investigate whether fine-tuning individual models based on questions of a similar nature (i.e., of the same skill category) improves performance for UA. At inference, predictions are made using the corresponding skill-based model. Secondly, we use these categories as the basis to create a custom development set for UQ, including questions which are representative of the skills and operators present in the UQ test set. We experiment between a manually chosen subset of 10 questions (n=1488), and a broader subset of 20 questions (n=1368), selected interactively with Claude Opus 4.6 (Anthropic, 2026a).

3.5 Ensemble Approach

Ensembles of multiple models have been successfully used for short answer scoring in several instances (Ormerod, 2022; Pathak et al., 2025). Majority voting of a number of individual models can help to mitigate limitations of single models, leading to more robust overall scoring. We utilize this majority vote approach in the UA tracks. To select an optimal combination, we evaluated a candidate pool of 100 checkpoints from our top experimental runs, based on QWK on the 3-way trial set. We performed a brute-force search over these candidates to identify the best combination of models for the ensemble. In line with Pathak et al. (2025), we use a combination of four models in the ensemble.

4 Results

We selected our test set submissions based on validation performance, using the provided trial set for unseen answers and custom train/validation splits for unseen questions. We aimed for a balanced representation of the strategies investigated during development. Official test results are in Table 1.

4.1 Unseen Answers

For the UA tracks, we tested fine-tuned Qwen3 (4B/8B) and Qwen3.5-9B models. During development, our best performing single model on the 3-way *trial* set was Qwen3-4B (0.792 QWK). Thus,

Rank	Submission ID	Base Model	Additional Details	QWK	Precision	Recall	F1
3-way Unseen Answers							
1	655003	Multi-Model	Ensemble: 1x9B, 2x8B, 1x4B	0.796	0.781	0.782	0.78
2	655652	Qwen3-8B	+FGM, +CL, +BG_correct, +full_data	0.79	0.772	0.77	0.771
8	648459	Qwen3-4B	*no special parameters*	0.781	0.779	0.767	0.77
9	654909	Qwen3-4B	by_skill	0.78	0.775	0.768	0.77
10	654389	Qwen3-8B	+CL, +BG_correct	0.779	0.771	0.763	0.765
2-way Unseen Answers							
1	655003	Multi-Model	Ensemble: 3→2way	0.726	0.887	0.887	0.887
4	654846	Multi-Model	Ensemble: Rank 7, 8, 10 submissions	0.71	0.88	0.881	0.881
7	654838	Qwen3-4B	by_skill, 3→2way	0.7	0.876	0.877	0.876
8	653676	Qwen3-4B	3→2way	0.698	0.875	0.877	0.876
10	653641	Qwen3-4B	2way	0.684	0.869	0.87	0.869
3-way Unseen Questions							
1	655694	Qwen3.5-9B	+NEFT, +CL, +BG_correct, +RR, custom_UQ_20	0.681	0.68	0.664	0.669
6	654994	Qwen3.5-9B	+NEFT, +CL, +BG_all, +RR, custom_UQ_20	0.635	0.657	0.64	0.644
13	653605	Qwen3.5-9B	+BG, +RR, +all_data, custom_UQ_10	0.591	0.727	0.601	0.6
16	654861	Qwen3.5-9B	+BG, +RR, custom_UQ_10	0.541	0.738	0.571	0.567
2-way Unseen Questions							
1	655727	Qwen3.5-9B	+NEFT, +CL, +BG_correct, +RR, custom_UQ_20, 3→2way	0.55	0.813	0.818	0.815
5	655638	Qwen3-4B	UA train, 2way	0.526	0.802	0.797	0.799
10	655654	Qwen3.5-9B	+NEFT, +CL, +BG_all, +RR, custom_UQ_20, 3→2way	0.501	0.796	0.804	0.797

Table 1: Test set results for IWM-DKM submissions for all four tracks. Abbreviations: BG = background knowledge (data source), CL = checklist thinking, RR= reframed rubrics, NEFT = NEFTune, by_skill = skill-specific classifiers, custom_UQ = custom validation set (with total questions), 3→2way = mapping 3-way label predictions to 2-way labels, UA train = trained only on Unseen Answers data, FGM = Fast Gradient Method adversarial training, full data = validation set included in training. Precision, Recall and F1 are all weighted metrics.

we submitted the 4B model, trained on either 3-way or 2-way labels, as baselines for the corresponding labeling scheme. We tested further strategies as described in Section 3, individually and in ensembles.

4.1.1 3-way Unseen Answers

The Qwen3-4B model achieved 8th place with a QWK on the test set of 0.781, providing a strong baseline. Our variations to this, such as fine-tuning individual Qwen3-4B models for each skill-based category (Rank 9), and including generated background knowledge and checklist thinking with the 8B model (Rank 10), performed on par, but were unable to beat the baseline. However, a variation of the Rank 10 submission that included trial data in training and was fine-tuned with the Fast Gradient Method (Miyato et al., 2017) demonstrated the strength of this form of adversarial training, placing second on the leaderboard. Our best-performing and final submission (0.796 QWK) was a majority-vote ensemble of four Qwen-based variations. Details of the models in this ensemble are shown in Appendix A. The ensemble was particularly effective at resolving ambiguous boundary cases between "Partially correct" and "Correct".

4.1.2 2-way Unseen Answers

We submitted the two variations of the baseline model (Qwen3-4B) to the 2-way track: firstly, mapping of the 3-way prediction labels to the 2-way labels (i.e., "Partially correct" is re-mapped to "Incorrect"); and secondly, fine-tuning directly on 2-way labels. Whilst both were strong baselines, the 2-way fine-tuned model performed marginally worse than the mapped version (0.684 QWK compared to 0.698 QWK). This suggested an advantage of making predictions at a more fine-grained level, thus, our remaining submissions for the 2-way UA track were mapped from 3-way labels. Mapping predictions from our by-skill Qwen3-4B classifiers led to marginal improvement over the baselines, and a majority-vote ensemble of the three systems ranked 7, 8 and 10 offered another slight performance improvement. However, the more architecturally diverse ensemble (mapped predictions from Rank 1 on the 3-way UA) resulted in our strongest submission, and the strongest overall submission for this track, achieving a QWK of 0.726, once again highlighting the robustness of ensemble setups.

4.2 Unseen Questions

4.2.1 3-way Unseen Questions

For the 3-way UQ track, we utilized Qwen3.5-9B as our base model, supplemented by checklist thinking, reframed rubrics, and background knowledge, and trained with a custom validation set. Submissions in Rank 1 and 6 were additionally fine-tuned with NEFTune (Jain et al., 2023). Between these submissions, we observe that silver

standard "Correct-only" student answers (Rank 1) provided a +0.046 QWK improvement over using all answers for Background Knowledge generation (Rank 6), underscoring that high-quality silver context is markedly more effective in driving generalization than a larger, noisier data pool, where misconceptions may be introduced. Furthermore, we experimented with two custom validation sets and saw a significant improvement in performance with the broader validation set (`custom_UQ_20`, used in Rank 1 and 6) compared to the 10 question set (`custom_UQ_10`, used for Rank 13 and 16[1]). Despite strong performance during development, the model optimized on the narrower, 10-question validation set showed poor transfer abilities to the test set, indicating the advantage of a broader pool of validation answers to aid reliable generalization.

4.2.2 2-way Unseen Questions

We applied several of our submissions from other tracks to the 2-way UQ track. To test generalizability from *unseen answers* to *unseen questions*, we evaluate Qwen3-4B fine-tuned on 2-way labels for UA on the 2-way UQ test set. Despite only training and validating on unseen answers, the model placed fifth on the leaderboard, although with a reduction in performance of 0.158 QWK (0.07 weighted F1). Additionally, we map the labels of our top two submissions from 3-way UQ, and reinforcing our earlier findings on the importance of high-quality background knowledge, the model trained with NEFTune, reframed rubrics, checklist thinking, the `custom_UQ_20` validation set, and background knowledge only outperformed the baseline when background knowledge was generated from silver-labels. On the first pass, where background knowledge was generated from all available student answers, overall performance was worse than our baseline of Qwen3-4B.

5 Discussion

Our results show that the outlined supplementary strategies improve fine-tuning performance. The results for both UQ tracks indicated the importance of *high-quality* additional information. Our leading submissions contained background knowledge generated from "Correct" student answers (as predicted by a first pass model), whereas results for an otherwise identical model with background knowledge based on all students answers performed significantly worse. We also observed during development that individual techniques taken alone can actually be *detrimental* to performance, their positive contribution only arises when combined.

We observed mixed results when it came to the optimization of the validation set for UQ. A broader 20-question trial set offered substantial improvements over a narrower 10-question trial set. Somewhat conflictingly, the Qwen3-4B model only optimized for unseen answers also exhibited promising performance in the 2-way UQ track, leading to further questions about the most impactful strategies to improve generalization in short answer grading.

Broadly, simple fine-tuning of decoder-only models was highly effective. Encoder-based models have generally dominated the short-answer scoring literature to date (e.g., Bexte et al., 2022; Sung et al., 2019), but despite experimenting with fine-tuning several German- and multilingual BERT-based models (e.g., ModernGBERT (Wunderle et al., 2025), gelectra-large (Chan et al., 2020)) in earlier trials, decoder-only architectures resulted in stronger performance. This reflects the NLP shift towards decoder-only architectures (Zhang et al., 2025; BehnamGhader et al., 2024), whose richer contextual embeddings better handle the nuanced semantics of rubric-based scoring. Finally, model ensembling proved beneficial, particularly in resolving boundary ambiguities between adjacent scoring categories. This is particularly advantageous in light of the fact that fine-tuned models exhibit sensitivity to small differences such as individual epochs and differences in seed. However, the high computational cost of this strategy limits the suitability of ensembles in real-life scoring.

6 Conclusion

Overall, our shared task submissions demonstrated the potential for fine-tuned decoder-only LLMs for rubric-based short answer scoring. Not only was LoRA fine-tuning of relatively small LLMs effective, but we highlighted how performance can be further improved by supplementing the input with task-relevant information. Our approach provides the model with more relevant information and presents that information in multiple structured formats to obtain more robust results. In future work, it would be interesting to further investigate the contributions of the individual strategies we used and to systematically evaluate their impact in other scoring contexts.

[1] A formatting mismatch in the reframed rubrics used in the Rank 16 submission may have also influenced performance.

Acknowledgments

This work was supported by the German Research Foundation (DFG) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the DFG through grant no. INST 37/1057-1 FUGG (Tübingen Machine Learning Cloud) and thank the Joachim Herz Stiftung for its support through the ALEE project.

Limitations

Due to the wide variety of methods we tested in the development phase, we were not able to systematically assess the impact of each individual technique on test set performance, nor did we evaluate all the models in our leading ensemble approaches as individual entries. Similarly, we did not test all our approaches on all tracks, which limits our ability to draw conclusions regarding the generalizability of methods across testing set-ups. Nevertheless, we believe our findings suggest support for the use of supervised fine-tuning of decoder-only models with LoRA in educational contexts. We used a single annotator for the skill-based categories. Double-annotation would strengthen the reliability of the categorization approach. In future work, it may be beneficial to test the techniques we implemented to improve upon standard fine-tuning on other datasets, and in more limited data scenarios, and to evaluate how expert knowledge could be used to further improve those techniques which were reliant on synthetic data (e.g., generated background knowledge, synthetically reframed rubrics).

Ethical Considerations

The use of AI for educational purposes has been designated as 'high risk' by the EU AI Act. All the data used in this shared task are anonymous and contain no personal identifying information, however, this remains an important consideration when thinking about the deployment of AI-based tools in real-world educational contexts (e.g., in classrooms). We acknowledge the use of AI tools for the generation of code in developing our systems, and for some of our data augmentation strategies, which carries both monetary and environmental costs.

References

Anthropic. 2026a. System card: Claude Opus 4.6. Technical report.

Anthropic. 2026b. System card: Claude Sonnet 4.6. Technical report.

Barbara, Ben Hamner, Jaison Morgan, lynnvandev, and Mark Shermis. 2012. The hewlett foundation: Short answer scoring. https://kaggle.com/competitions/asap-sas. Kaggle.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling, COLM 2024*.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring - how to make S-BERT keep up with BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 118–123, Seattle, Washington. Association for Computational Linguistics.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th international conference on computational linguistics*, pages 6788–6796.

Jonathan Cook, Tim Rocktäschel, Jakob Foerster, Dennis Aumiller, and Alex Wang. 2024. Ticking all the boxes: Generated checklists improve llm evaluation and generation. *Preprint*, arXiv:2410.03608.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Scott Frohn, Tyler Burleigh, and Jing Chen. 2025. Automated scoring of short answer questions with large language models: Impacts of model, item, and rubric design. In *Artificial Intelligence in Education*, pages 44–51, Cham. Springer Nature Switzerland.

Xunyi Gao, Shamya Karumbaiah, Adi Dalal, Indrani Dey, Dana Gnesdilow, and Sadhana Puntambekar. 2025. A comparative analysis of llm and specialized nlp system for automated assessment of science content. In *Artificial Intelligence in Education*, pages 76–82, Cham. Springer Nature Switzerland.

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.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, and 1 others. 2023. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*.

Lan Jiang and Nigel Bosch. 2024. Short answer scoring with GPT-4. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pages 438–442, Atlanta GA USA. ACM.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

Michael Pin-Chuan Lin, Jing-Yuan Huang, Daniel H. Chang, Gerald Tembrevilla, G. Michael Bowen, Eric Poitras, Vasudevan Janarthanan, and Jeeho Ryoo. 2026. Open-source large language models in education: A narrative review of evidence, pedagogical roles, and learning outcomes. *AI in Education*, 2(1).

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, Scottland, UK. Association for Computational Linguistics.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations (ICLR)*.

Christopher Ormerod. 2022. Short-answer scoring with ensembles of pretrained language models. *arXiv preprint*. Version Number: 1.

Aditya Pathak, Rachit Gandhi, Vaibhav Uttam, Arnav Ramamoorthy, Pratyush Ghosh, Aaryan Raj Jindal, Shreyash Verma, Aditya Mittal, Aashna Ased, Chirag Khatri, Yashwanth Nakka, Devansh, Jagat Sesh Challa, and Dhruv Kumar. 2025. Rubric is all you need: Improving llm-based code evaluation with question-specific rubrics. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V.1*, ICER '25, page 181–195, New York, NY, USA. Association for Computing Machinery.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267. Version 5.1.

Chul Sung, Tejas I. Dhamecha, and Nirmal K. Mukhi. 2019. Improving Short Answer Grading Using Transformer-Based Pre-training. In *International Conference on Artificial Intelligence in Education*.

Tianqi Wang, Naoya Inoue, Hiroki Ouchi, Tomoya Mizumoto, and Kentaro Inui. 2019. Inject rubrics into short answer grading system. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for*

Low-Resource NLP (DeepLo 2019), pages 175–182, Hong Kong, China. Association for Computational Linguistics.

Julia Wunderle, Anton Ehrmanntraut, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. New encoders for german trained from scratch: Comparing moderngbert with converted llm2vec models. *arXiv preprint arXiv:2505.13136*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: self-taught reasoner bootstrapping reasoning with reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Biao Zhang, Yong Cheng, Siamak Shakeri, Xinyi Wang, Min Ma, and Orhan Firat. 2025. Encoder-decoder or decoder-only? revisiting encoder-decoder large language model. *Preprint*, arXiv:2510.26622.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey. *ArXiv*, abs/2402.19473.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

A Ensemble model details

We list the details of the models included in our four-model ensemble for 3-way unseen answers submission in Table 2.

Base model	Details
Qwen3.5-9B	+BG_correct +CL
Qwen3-8B	+FGM, +BG_correct, +CL
Qwen3-8B	+DPO
Qwen3-4B	+BG_correct, +CL

Table 2: Models included in ensemble for Rank 1 Unseen Answers 3-way submission (Abbreviations: BG_correct = Background Knowledge from correct answers, CL = checklist thinking, FGM = Fast Gradient Method, DPO = Direct Preference Optimization)

B Example Item with Background Knowledge and Reframed Rubric

Table 3 shows an example item from the ALICE-LP dataset, illustrating the question, generated background knowledge, original scoring rubric, reframed rubric, operator, and skill category. The background knowledge field contains the generated passage used as supplementary material during fine-tuning (Section 3.3). This item is representative of the *Concept Description* skill category, with the operator *Beschreiben ('Describe')* (Section 3.4).

C Prompts

C.1 Prompt for Reframed Rubrics

The prompt used with GPT5.1 to generate reformulations of the 3-way rubrics in decision tree format is shown in Listing 1.

C.2 Prompt for Background Knowledge

The prompt used with Claude Sonnet for generating question-specific background knowledge is provided in Listing 2.

D Skill-based categories

Table 4 shows an example question for each of the four categories we annotate the data with, as described in (Section 3.4).

Category	Example
Question	Beschreibe das Power-to-Gas Verfahren in eigenen Worten. *'Describe the power-to-gas process in your own words.'*
Background Knowledge	Beim Power-to-Gas-Verfahren wird überschüssige elektrische Energie, die zum Beispiel durch Windkraftanlagen erzeugt wird, in chemische Energie umgewandelt und so langfristig gespeichert. Dazu durchläuft die Energie eine Kette von Umwandlungsschritten: Zunächst wird Bewegungsenergie des Windes im Windrad in elektrische Energie umgewandelt, die dann einem Elektrolyseur zugeführt wird. Im Elektrolyseur wird mithilfe der elektrischen Energie Wasser in seine Bestandteile Wasserstoff und Sauerstoff aufgespalten, wobei die Energie in Form von chemischer Energie im Wasserstoff gespeichert wird. Der erzeugte Wasserstoff kann anschließend in Gasspeichern oder im Erdgasnetz gelagert werden, bis er bei Bedarf in einem Gaskraftwerk wieder in elektrische Energie umgewandelt wird. Die vollständige Energieumwandlungskette lautet also: Bewegungsenergie → elektrische Energie → chemische Energie (Wasserstoff) → elektrische Energie. *'In the power-to-gas process, excess electrical energy—generated, for example, by wind turbines—is converted into chemical energy and thus stored for the long term. To achieve this, the energy undergoes a series of conversion steps: First, the kinetic energy of the wind is converted into electrical energy in the wind turbine, which is then fed into an electrolyzer. In the electrolyzer, electrical energy is used to split water into its components, hydrogen and oxygen, with the energy being stored in the hydrogen in the form of chemical energy. The hydrogen produced can then be stored in gas storage facilities or in the natural gas grid until it is converted back into electrical energy in a gas-fired power plant when needed. The complete energy conversion chain is therefore: kinetic energy → electrical energy → chemical energy (hydrogen) → electrical energy.'*
Rubric	(Incorrect) Die SuS formulieren keine korrekte Energieumwandlungskette. *'The students do not formulate a correct energy conversion chain.'* (Partially correct) Die SuS formulieren eine teils korrekte Energieumwandlungskette (mindestens die Hälfte) *'The students formulate a partially correct energy conversion chain (at least half of it)'* (Correct) Die SuS formulieren die vollständige Energieumwandlungskette. *'The students describe the complete energy conversion chain.'*
Reframed Rubric	1. Beschreibt die Antwort überhaupt eine Energie-Umwandlungskette beim Power-to-Gas-Verfahren? *'Does the answer even describe an energy conversion chain in the power-to-gas process?*' - **Nein** → **Incorrect** , **Ja**→ 2 2. Wird deutlich, dass elektrische Energie (z.B. aus erneuerbaren Quellen) zunächst in chemische Energie von Wasserstoff umgewandelt wird (Elektrolyse von Wasser)? *'Is it clear that electrical energy (e.g., from renewable sources) is first converted into the chemical energy of hydrogen (through the electrolysis of water)?'* - **Nein**→ **Incorrect**, **Ja** → 3 3. Wird zusätzlich beschrieben, dass aus dem Wasserstoff ein gasförmiger Energieträger (z.B. Methan) entsteht, der gespeichert / ins Gasnetz eingespeist werden kann? *'Is it also described that hydrogen is converted into a gaseous energy carrier (e.g., methane) that can be stored or fed into the gas grid?'* - **Nein** → - Wenn nur ein Teil der Umwandlungskette korrekt ist (z.B. Strom→ Wasserstoff, aber ohne weitere Verwendung / Methanisierung) *'If only part of the conversion chain is correct (e.g., electricity → hydrogen, but without further use or methanation)'* → **Partially correct** - **Ja** → 4 4. Ist die gesamte Energieumwandlungskette inhaltlich richtig und in der Reihenfolge vollständig beschrieben (z.B. erneuerbare elektrische Energie→ Elektrolyse→ Wasserstoff→ ggf. Methanisierung→ Speicherung / Nutzung im Gasnetz)? *'Is the entire energy conversion chain described accurately contentwise and completely in terms of sequence (e.g., renewable electrical energy → electrolysis → hydrogen → methanation, if applicable → storage/use in the gas grid)?'* - **Nein** → **Partially correct**, **Ja** → **Correct**
Operator	Beschreiben (*'describe'*)
Skill Category	Concept Description
Example Student Answer	Die Elektrische Energie wird in Chemische umgewandelt welche dann wieder zu Elektrischer wird. *'Electrical energy is converted into chemical energy, which is then converted back into electrical.'*
Gold Label	Partially correct

Table 3: Example item from the ALICE-LP dataset. The example student answer correctly identifies two of the three conversion steps (electrical → chemical → electrical) but omits the initial step (kinetic energy → electrical energy via wind turbine), warranting a *Partially correct* label. (Original is in German; English translation in quotes added here for transparency.)

```
Sie sind pädagogische/r Assistent/in und haben die Aufgabe, leicht verständliche Bewertungsrubriken f
    ür Fragen in formativen Schülerbeurteilungen zu entwerfen. \ Der Bildungskontext sind deutsche
    Gemeinschaftsschulen und Gymnasien und die abgedeckten Fachbereiche sind: Biologie, Chemie,
    Physik und Mathematik.
‘You are a pedagogical assistant, whose goal is to help design easy-to-follow grading rubrics to
    questions in formative student assessments. \ The educational setting is German
    Gemeinschaftsschulen and Gymnasien and the subject domains covered are: Biology, Chemistry,
    Physics and Mathematics.’
Ihnen wird eine Frage und eine bestehende Rubrik mit drei Kategorien ‘Incorrect’, ‘Partially correct’
     und ‘Correct’ vorgelegt. \ Ihre Aufgabe besteht darin, die Rubrik im Format eines
    Entscheidungsbaums mit Fragen neu zu schreiben, um die richtige Note zuverlässiger zu vergeben.
    Stellen Sie sicher, dass alle relevanten Informationen aus der bestehenden Rubrik in den
    Entscheidungsbaum aufgenommen werden.
‘You will be presented with a question, and an existing rubric containing three categorical levels ‘
    Incorrect’, ‘Partially correct’ and ‘Correct’. \ Your task is to re-write the rubric in the
    format of a decision tree of questions to assign the correct grade more reliably. Make sure all
    the relevant information from the existing rubric is included in the decision tree.’
Hier ist die Frage: {question}
‘Here is the question: {question}’
   ***
Hier ist die bestehende Rubrik.
‘Here is the existing rubric.’
{rubric}
   ***
Neu gestaltete Entscheidungsbaum-Benotungshilfe:
‘Newly designed decision tree grading aid:’
```

Listing 1: Prompt used to generate reframed rubrics with GPT5.1 (Original is in German; English translation in quotes added here for transparency.)

```
Schreibe einen Lernzettel-Eintrag (3-5 Saetze Fliesstext, kein Markdown) zu dem
    naturwissenschaftlichen Thema, das in der folgenden Aufgabe behandelt wird. Der Lernzettel soll
    die wichtigsten Fachbegriffe, Konzepte und Zusammenhaenge erklaeren, die man kennen muss, um die
     Aufgabe vollstaendig beantworten zu koennen. Schreibe so, als wuerdest du einem Mitschueler das
     noetige Vorwissen kompakt zusammenfassen.
‘Write a study note entry (3-5 sentences of continuous text, no Markdown) on the scientific topic
    covered in the following task. The study note should explain the most important technical terms,
     concepts, and relationships that one needs to know in order to fully answer the task. Write as
    if you were compactly summarizing the necessary prior knowledge for a classmate.’

Aufgabe: {question}
‘Task: {question}’

Zur Orientierung, welche Konzepte relevant sind:
‘For orientation on which concepts are relevant:’
- Incorrect: {rubric_incorrect}
- Partially correct: {rubric_partial}
- Correct: {rubric_correct}

Korrekte Schülerantworten ({n} Beispiele):
‘Correct student answers ({n} examples):’
- {Student answer 1}
- {Student answer 2}
{...}
- {Student answer n}
```

Listing 2: Prompt used to generate background knowledge with Claude Sonnet (Original is in German; English translation in quotes added here for transparency.)

Category	Example
Concept Description	Beschreibe auf Grundlage Deiner oben getroffenen Auswahl in eigenen Worten, was Du allgemein unter der Reaktionsgeschwindigkeit chemischer Reaktionen verstehst. *'Based on the selection you made above, describe in your own words what you generally understand by the reaction rate of chemical reactions'*
Example Led	Beurteile das Zerteilen der Äpfel hinsichtlich der Geschwindigkeit des Verderblichkeitsprozesses (Hinweis: wenn Du fachliche Informationen zu chemischen Abläufen des Verderbens einbeziehen möchtest, kannst Du diese unter Angabe der Quellen recherchieren und einbinden). *'Evaluate how cutting apples affects the rate at which they spoil (Note: If you wish to include technical information on the chemical processes involved in spoilage, you may research and incorporate this information, provided you cite your sources).'*
Idea Seeking	Hast Du auf Grundlage der oben beantworteten Aufgaben bereits Ideen, welche Größen geeignet sein könnten, um die Geschwindigkeit dieser Reaktion zu erfassen? Wenn ja, beschreibe und begründe Deine Ideen. *'Based on your answers to the questions above, do you already have any ideas about which quantities might be suitable for measuring the rate of this reaction? If so, describe and explain your ideas.'*
Mathematical	In 2.2 Untersuchung der Lampen 4 habt ihr die Formel zur Berechnung der elektrischen Leistung genannt. Berechnet mit Hilfe der Formel die elektrischen Leistungen der einzelnen Lampen. *'In Section 2.2, "Examination of the Lamps 4," you were given the formula for calculating electrical power. Use this formula to calculate the electrical power of each lamp.'*

Table 4: Examples of questions for each skill-based category (Original in German; English translation in quotes added here for transparency)