Report on the BEA 2026 Shared Task on Rubric-based Short Answer Scoring for German

Sebastian Gombert[1], **Zhifan Sun**[1], **Fabian Zehner**[1,2], **Jannik Lossjew**[3], **Tobias Wyrwich**[4], **Berrit K. Czinczel**[4], **David Bednorz**[4], **Sascha Bernholt**[4], **Knut Neumann**[4], **Ute Harms**[4], **Aiso Heinze**[4], **Hendrik Drachsler**[1,5,6]

[1]DIPF | Leibniz-Institute for Research and Information in Education
[2]Centre for International Student Assessment (ZIB)
[3]Institute for Chemistry Education, Eberhard Karls University of Tübingen
[4]IPN | Leibniz Institute for Science and Mathematics Education
[5]Studiumdigitale & [6]Computer Science Department, Goethe University Frankfurt

Abstract

We present the BEA 2026 shared task on rubric-based short answer scoring for German. Rubric-based short answer scoring is a case of automatic short answer scoring (ASAS) that requires models to apply textual scoring rubrics to student answers as a basis for assigning scores. For the shared task, we introduced a novel German-language dataset from multiple STEM domains to provide a comprehensive benchmark for this problem. The dataset was designed to evaluate both performance and generalization (the latter, by distinguishing between seen and unseen questions), as well as coarse- and fine-grained scoring (2-way vs. 3-way). The systems submitted to the shared task cover a wide range of approaches, including fine-tuned large language models, prompt-based methods, human–AI collaboration strategies, or a combination of these. The results show that structured, task-adapted LLM systems achieved the strongest performance across all tracks. The winning system, IWM-DKM, combined LoRA fine-tuning of Qwen models with rubric-aware input structuring, including checklist-style reasoning, rubric reframing as decision trees, background knowledge injection, and ensemble voting. Other systems similarly relied on fine-tuned LLMs, retrieval-augmented prompting, encoder–LLM ensembles, or weighted aggregation strategies. Overall, the shared task results show that rubric-based scoring benefits most from systems that explicitly operationalise rubric semantics, while generalisation to unseen questions remains a central challenge.

1 Introduction

Short answer scoring is an established task in educational natural language processing (Burrows et al., 2015; Bai and Stede, 2023; Bexte et al., 2024). It was already addressed in at least two previous shared tasks, namely the *SemEval 2013 Shared Task* (Dzikovska et al., 2013) and the *ASAP-SAS* Kaggle competition (Hamner et al., 2012). In recent years, it has primarily been approached using two paradigms: instance-based and similarity-based approaches (Bexte et al., 2022, 2023).

On the one hand, instance-based approaches train a model that predicts scores as labels or regression targets for a given input response. While this approach has been shown to work quite well, it also suffers from the problem of not transferring well to questions unseen during training (Padó et al., 2023). This means that implementing an instance-based short-answer scoring system requires collecting and manually scoring student answers for a select set of questions. A model trained on these in an instance-based fashion then only supports scoring unseen answers to these known questions. In this approach, the resulting model is intended to implicitly take up the relevant rules for scoring instances.

Similarity-based approaches, on the other hand, attempt to circumvent these problems by selecting a setup based on nearest-neighbour search or natural language inference (Bexte et al., 2022; Camus and Filighera, 2020), which are designed to compare a student's answer to one or more provided reference answers. While this results in better transfer capabilities to unseen questions and domains, a problem with this setup is that not all possible question formats necessarily allow using a similarity-based setup, since for some questions, a broad range of answers might be valid, and out-of-sample generalization may fail due to unexpected content during inference.

With this shared task, we therefore look at another interpretation of the task that could help to circumvent these problems: rubric-based short answer scoring. For rubric-based short answer scoring, a model is provided with an answer, a question/prompt and a textual scoring rubric that, for each possible score, establishes criteria that need to be met so the same can be assigned to a given answer. To successfully solve this task, a model must

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1179–1192

account for the semantics of a scoring rubric and apply its criteria to the candidate answers. This setup is inspired by established human scoring practice (Reddy and Andrade, 2010; Panadero and Jonsson, 2013). Here, human scorers are often provided with a rubric and then need to reason about which criteria apply to a given answer to assign the appropriate score. These rubrics are often question-specific and are, in turn, typically embedded in assessment frameworks that specify the construct that the test is intended to measure (e.g. OECD, 2024, 2023).

While rubrics have been applied in various works on free-text scoring and have shown promise as supplementary input or pre-training resource for instance- and similarity-based scoring systems (Wang et al., 2019; Li et al., 2023; Condor et al., 2022; Sonkar et al., 2024) as well as for systems based on LLM prompting (Wei et al., 2025; Frohn et al., 2025; Cong et al., 2026), there is, as of now, a lack of focused research on rubric-based short-answer scoring. However, this involves many challenges such as the uncertainty in the rubrics and, depending on the question, the semantic variability of the answers. With additional information from rubrics or assessment frameworks, resulting models can be designed more explicitly to incorporate scoring criteria that are specific to a given domain or assessment construct. As of now, there is a clear research gap in understanding how well current methods in NLP can comprehend rubrics and apply them to students' answers to predict corresponding scores. With this shared task, we aimed to kickstart research on this specific interpretation of the short-answer scoring task with the following concrete contributions:

- We hosted a shared task on *rubric-based short answer scoring*, a setting that reflects real-world educational assessment, ranging from classroom to standardized testing, and requires models to interpret and apply textual scoring criteria.
- We provided a new publicly available benchmark dataset for this task in German, contributing to the currently limited pool of non-English resources for ASAS.
- We designed a comprehensive evaluation setup that systematically distinguishes between seen and unseen questions as well as coarse- and fine-grained scoring (2-way vs. 3-way), enabling a detailed analysis of generalization capabilities.
- In this paper, we present a comparative overview of the submitted modeling approaches, highlighting emerging design patterns such as structured LLM reasoning, prompt optimization, hybrid modeling, and human–AI collaboration.

2 Related Work

Automatic short answer scoring (ASAS) is an established task in educational NLP, where systems aim to assign scores to short written student responses (Burrows et al., 2015; Bai and Stede, 2023; Bexte et al., 2024). Recent work has mainly followed two underlying paradigms (Bexte et al., 2023). Instance-based approaches treat scoring as text classification or regression, where student answers are mapped directly to score labels (Riordan et al., 2017; Gombert et al., 2023). These methods work well for questions seen during training, but typically transfer less reliably to unseen questions (Padó et al., 2023), especially when the new questions assume a different number of levels. Similarity-based approaches instead compare student answers to reference answers or representative answer clusters, for example, using natural language inference or sentence embeddings (Bexte et al., 2022, 2023; Zehner et al., 2016). While this can improve reuse across questions, such methods depend on the availability of reference answers at all available levels and full semantic coverage is not always guaranteed.

Rubrics are central to human scoring practice because they define explicit criteria for assigning performance levels (Reddy and Andrade, 2010; Panadero and Jonsson, 2013). In automatic scoring, however, they have mostly been used as additional input rather than as the main scoring reference. Earlier work has injected rubric information into neural scoring models (Wang et al., 2019) or used rubrics for further pre-training (Condor et al., 2022). These studies suggest that rubrics can improve scoring, especially in low-resource and transfer settings, but focused research on rubric-based short answer scoring remains limited.

Recent work has started modelling rubrics as central semantic structures for scoring. Gombert et al. (2026b) introduced the *GRAASP* and *ToLeGRAA* architectures, in which scoring is modelled as the alignment between student responses and

rubrics, using attention over transformer-based representations. Building on this direction, the *AGRAA* framework (Gombert et al., 2026a) models rubric descriptors as low-dimensional semantic subspaces derived from contextualised embeddings, so that scoring corresponds to measuring how strongly a response aligns with rubric-induced aspect spaces.

In parallel to model-based approaches, recent work has explored *rubric-based prompting* of large language models (LLMs) for automated assessment. Instead of fine-tuning, these approaches provide rubric criteria directly within the prompt and ask the model to assign scores based on them. Cong et al. (2026) showed that incorporating rubric descriptors into prompts substantially improves scoring performance compared to prompting without rubrics, highlighting that rubrics provide more structured guidance than standard few-shot examples. Similarly, Wei et al. (2025) demonstrate that *concept-based rubrics* not only improve LLM scoring performance but also enable the generation of synthetic training data for downstream supervised models. Overall, these findings suggest that even without task-specific training, LLMs can leverage rubric structures when they are explicitly encoded in prompts, although their performance and reliability remain sensitive to rubric design and task complexity.

3 Dataset: ALICE-LP 1.1

The dataset we used in this shared task is a novel dataset called *ALICE-LP 1.1*, which contains mid- and high-school-level answers to questions from four STEM domains: *physics* (Wyrwich et al., 2025), *mathematics* (Bednorz et al., 2024), biology (Czinczel et al., 2025), and chemistry (Lossjew and Bernholt, 2025). These answers were collected within the context of the *ALICE* project, funded by the Leibniz Foundation, at middle and high schools in the German state of Schleswig-Holstein. The main goal of this project was to conduct learning progression analytics (Kubsch et al., 2022) in the classroom. Teachers guided students through lessons which involved solving various interactive learning activities in Moodle courses synchronously. This included a set of short-answer items, which are the basis of this dataset.

ALICE-LP 1.1 is a German ASAS dataset. Each question is rated on a three-point scale (incorrect, partially correct, correct). Each datapoint consists of the question text, the answer, a textual two- as well as three-way rubric defining the criteria for each possible score, and the corresponding score. Table 2 shows an examplary datapoint from this dataset. The two-way rubric has been artificially created by merging the partially correct and incorrect score levels.

Set	#Questions	#Answers	#Levels (0/1/2)
Train	78	7,899	2,728/2,863/2,308
UA	78	2,008	691/726/591
UQ	39	3,086	1,146/1,021/919

Table 1: Distributional properties of *ALICE-LP 1.1*.

The answers were scored by multiple student assistants enrolled in teacher education programmes for the respective subjects. For each question, an agreement was measured using a subset of answers. Student assistants were trained to score each question using a subset of answers until an agreement of at least $\kappa \geq .65$ was reached. The remaining answers were then distributed among the raters. The training set includes a total of 7,899 answers. The evaluation set is divided into two subsets: *unseen answers* and *unseen questions*. This setup was inspired by the established SemEval-2013 (Dzikovska et al., 2013) and SAF (Filighera et al., 2022) datasets, enabling evaluation of how well systems transfer to unseen questions. The *unseen answers* dataset includes 2,008 answers to questions contained in the training set (using a question-wise stratified 80/20 split). In contrast, the *unseen questions* dataset contains 3,086 answers to questions not in the training set. An earlier version of this dataset (*ALICE-LP 1.0)* was used in Gombert et al. (2026b) and Gombert et al. (2026a). For the shared task and the published benchmark dataset, we removed one question with 82 responses from the unseen questions dataset because we found it requires a reference answer to be properly gradable. Table 1 shows the distributional properties of *ALICE-LP 1.1*.

4 Shared Task Structure

In the context of the shared task, we hosted four evaluation tracks. All tracks were evaluated using the same set of evaluation metrics: weighted precision, weighted recall, weighted F1, and quadratic weighted kappa (QWK). QWK served as the primary metric for leaderboard ranking for each track.

Unseen Answers 2-Way For this track, participating models were evaluated based on how well

Input Category	Example
Question	Name consequences that the gas shortage could have for Germany and your school.
Answer	The school might have to close because it can no longer be heated. It could also be that students just have to wear jackets during lessons.
Rubric	(2) Students identify at least two links between a gas supply stop and the supply of electricity and/or heating. (1) Students identify one link between a gas supply stop and the supply of electricity and/or heating. (0) Students do not identify a link between a gas supply stop and the supply of electricity and/or heating.
Score	0/1 (2-Way) 1/2 (3-Way)

Table 2: An example question with the corresponding rubric, one example student answer and the corresponding score taken from the *ALICE-LP* dataset (translated from German to English).

they could score unseen answers to questions seen during training. For this track, we distinguished only between correct and incorrect answers, with partially correct answers counted as incorrect.

Unseen Questions 2-Way For this track, participating models were evaluated based on how well they could score answers to questions not seen during training. For this track, we distinguished only between correct and incorrect answers, with partially correct answers counted as incorrect.

Unseen Answers 3-Way For this track, participating models were evaluated based on how well they could score unseen answers to questions seen during training. For this track, we distinguished between correct, partially correct, and incorrect answers.

Unseen Questions 3-Way For this track, participating models were evaluated based on how well they could score answers to questions not seen during training. For this track, we distinguished between correct, partially correct, and incorrect answers.

All four tracks were hosted on Codabench (Xu et al., 2022). Participants were allowed to submit up to ten submissions per track, with the best submission, as determined by QWK, counted towards the official leaderboard of the shared task. We allowed multiple submissions to account for different variants of authors' systems and the non-deterministic nature of deep learning methods. In the Appendix, we provide tables of the five best runs per team and track.

5 Submissions

Except for the teams *Diffuser* and *Afrilan*, all participating teams submitted system description papers. We summarise the results in the following section. All teams participated in the *Unseen Answers 2-way* track, making it the most complete comparison setting. For the *Unseen Questions 2-way* track, all teams except *Diffuser* submitted results. In the *Unseen Answers 3-way* track, again all teams except *HFT* participated, while the *Unseen Questions 3-way* track saw the lowest participation, with only *IWM-DKM*, *WSE Research*, *SDPA*, *ASLAN*, *Afrilan*, and *AMATI* submitting results. Overall, participation decreases slightly in more challenging and fine-grained settings, particularly in the unseen-question and 3-way-classification scenarios.

Most participating teams were affiliated with German institutions, which is expected, given that the dataset and task are centred on German-language student answers. At the same time, the shared task attracted participants from outside Germany, indicating broader international interest in rubric-based short-answer scoring. This suggests that, despite the language-specific setting, the underlying problem formulation is relevant beyond the German research community.

5.1 System Descriptions

While this section systematises the submissions' methodology, Table 3 provides a brief summary.

IWM-DKM (Belcher et al., 2026) The approach is based on supervised fine-tuning of decoder-only LLMs (Qwen family) with LoRA, where the model jointly processes the question, student answer, and rubric to predict scores. Performance is improved by systematically enriching and structuring the input: (i) checklist thinking introduces intermediate yes/no decisions aligned with rubric levels to guide reasoning, (ii) rubric reframing converts rubrics into structured decision trees for clearer decision

Team	Keywords of the Approach	Models / LLMs (Final System)
IWM-DKM (Belcher et al., 2026)	LoRA fine-tuning + (checklist thinking, rubric reframing, background knowledge injection); ensemble voting	Qwen-3-4B (QLoRA), Qwen-3.5-9B (QLoRA)
WSE Research (Gwozdz and Both, 2026)	LoRA fine-tuning + (instruction-following format, 7B–72B scaling); rubric-conditioned prompting + (TF-IDF exemplar retrieval); ensemble aggregation + (QWK-weighted voting)	Qwen-2.5-72B (QLoRA), Gemini 3 Flash + (Prompting + tf-idf-based RAG), kNN with tf-idf
SDPA (Liu and Zhang, 2026)	LoRA/DoRA fine-tuning + (Alpaca instruction format, direct label prediction); ensemble + (trained encoder, few-shot Claude, optimised soft voting)	GELECTRA-large, XLM-RoBERTa-large, BERT-base-german-cased, Claude (few-shot), Llama 3.1 8B (LoRA/DoRA)
ASLAN (Bexte et al., 2026)	Similarity-based scoring + (LaSiLearn, SetFit); instance-based classification + (domain-specific fine-tuning); zero-shot rubric-based prompting; majority voting	G-SciEdBERT, LaSiLearn / SetFit + (mpnet, mini-lm, distil), Phi-4, Gemma 3, GPT-5-mini + (ensemble via majority voting)
AMATI (Willis and Third, 2026)	Inductive Logic Programming + (ASP-based rule induction, lemma matching, key information rules); rule-guided prompting + (ILP theory injection, in-context training examples)	Mistral combined with ILP-derived rules
RETUYT-INCO (Sastre et al., 2026)	Meta-prompting + (group-specific prompt generation, per-group selection); prompt tuning + (per-group soft prompt embeddings); synthetic data augmentation + (class balancing, noise injection)	Gemini 3 Flash
HFT (Padó, 2026)	Zero-shot rubric-based prompting + (role prompting, expert committee prompting); hybrid human–AI grading + (confidence-based routing, pre-defined quality threshold)	GPT-OSS-120B

Table 3: Overview of final systems submitted by participating teams, focusing on core modelling approaches and models used for official submissions.

boundaries, and (iii) background knowledge injection provides generated domain context to support understanding. These techniques are complemented by strategies such as skill-based data partitioning and tailored validation setups, and are ultimately combined in ensemble models using majority voting to increase robustness, particularly for ambiguous cases, with gains arising from the interaction of these components rather than any single method alone.

WSE Research (Gwozdz and Both, 2026) The approach combines three complementary scoring strategies within a unified pipeline: rubric-based prompting, LoRA-based fine-tuning of Qwen models, and prediction aggregation. Given the input triple (question, student answer, rubric), candidate scores are generated via prompted LLMs (optionally enhanced with TF-IDF–based example retrieval), multiple fine-tuned models across different scales, and a non-neural TF-IDF kNN baseline. Fine-tuning is performed in an instruction-following format that mirrors inference, enabling models to internalise rubric criteria more effectively. Scaling experiments show that performance improves with model size but saturates at larger scales. Predictions are combined using weighted voting or stacking, where aggregation improves performance on seen questions by leveraging complementary error patterns, whereas for unseen questions, a single large fine-tuned model generalises best.

SDPA (Liu and Zhang, 2026) The approach employs parameter-efficient fine-tuning of LLMs for rubric-based short answer scoring under low-resource conditions, alongside an ensemble baseline. It compares two complementary paradigms: (i) an encoder-based ensemble that combines multiple fine-tuned transformer models with few-shot LLM predictions (e.g., Claude) via weighted soft voting, and (ii) instruction-tuned LLMs (Llama3.1-8B) adapted using PEFT methods such as LoRA and DoRA. The latter directly predicts rubric-aligned labels from structured inputs (question, answer, rubric), enabling the model to internalise rubric reasoning rather than relying on prompt-based inference. Results show that while ensem-

bles benefit from combining complementary signals, particularly on seen questions, parameter-efficient fine-tuned LLMs consistently outperform both prompt-based and ensemble approaches, especially for unseen questions where generalisation to new rubrics is critical.

ASLAN (Bexte et al., 2026) The approach combines three complementary scoring paradigms: similarity-based methods, instance-based classification, and rubric-prompted LLMs, and evaluates them across seen and unseen settings. Similarity-based models (e.g., SBERT variants with LaSiLearn or SetFit) compare student answers to labelled training answers, while instance-based models (e.g., BERT variants) learn direct mappings from input text to scores. For unseen questions, where no task-specific training data is available, zero-shot LLMs are prompted with the rubric to assign labels. Across all paradigms, predictions are combined using majority voting, leveraging their complementary strengths. Results show that similarity-based approaches perform best when training data for a question is available, while LLM-based methods are more suitable for generalising to unseen questions. Ensemble methods consistently improve robustness and overall performance by integrating signals from different scoring strategies.

AMATI (Willis and Third, 2026) The approach combines symbolic rule learning with large language models to balance interpretability and robustness. First, Inductive Logic Programming (ILP) is used to induce human-readable grading rules from training data, where questions, rubrics, and student answers are represented as logical predicates and optimised via Answer Set Programming to generate rule-based scoring theories. These rules capture lexical and structural patterns aligned with rubric criteria, enabling transparent and explainable predictions. To address the brittleness of symbolic rules, the system integrates them with an LLM (Mistral) by incorporating either training examples or induced rules into the prompt. While the LLM alone provides robust scoring, especially for noisy student language, combining it with induced rules improves performance in the more fine-grained 3-way setting.

RETUYT-INCO (Sastre et al., 2026) The approach centres on meta-prompting, where an LLM is used to automatically generate task-specific scoring prompts tailored to each question–rubric pair. In an offline phase, the model analyses the rubric and labelled training answers for a given group to produce a reusable, optimised prompt that captures the scoring criteria more precisely than a generic template. At inference time, new student answers are routed to their corresponding group and evaluated using the generated prompt, effectively adapting the scoring behaviour to the specific characteristics of each task. To further improve performance, the method incorporates prompt variation, selection based on validation performance, and synthetic data generation to address class imbalance.

HFT (Padó, 2026) The approach explores a low-cost, blunt-edge grading setup based on zero-shot prompting of open-source LLMs, combined with a hybrid human–machine workflow. The system uses simple rubric-grounded prompts with off-the-shelf LLMs to produce initial grade predictions, prioritising robustness, reproducibility, and applicability in resource-constrained settings. While these models perform moderately well, especially in information-poor scenarios, they are not entirely competitive with fine-tuning-based approaches when richer task-specific data is available. To compensate for this, the method incorporates human review in a structured hybrid grading process: model predictions for the most reliable class are accepted automatically, while less certain cases are manually reviewed.

6 Results

Table 4 shows the best submitted result per team and track. Overall, the results reflect the methodological differences discussed above and reveal a consistent hierarchy across settings for the shared task's dataset, *ALICE-LP 1.1*. In the information-specific Unseen Answers conditions that allow training question-specific models, strongly optimised, model-centric approaches dominate: *IWM-DKM* and *WSE Research* consistently achieve the top ranks across both 2-way and 3-way tasks, with *SDPA* closely following. This confirms that structured fine-tuning, input engineering, and ensembling are highly effective when scored answers are available for training. In contrast, more prompt-centric approaches (sensu instruction learning) such as *RETUYT-INCO* and multi-paradigm ensembles such as *ASLAN* achieve competitive but slightly lower performance, indicating that while prompt optimisation and multi-paradigm aggregation are beneficial, they do not fully match the

Rank	Team Name	↑ Quadratic Weighted Kappa	Weighted Precision	Weighted Recall	Weighted F1-Score
		Unseen Answers 2-way			
1	IWM-DKM	**.726**	**.887**	**.887**	**.887**
2	WSE Research	.717	.882	.883	.883
3	SDPA	.682	.869	.866	.867
4	RETUYT-INCO	.674	.865	.861	.863
5	ASLAN	.663	.861	.864	.862
6	AMATI	.644	.857	.843	.847
7	Diffuser	.615	.840	.842	.841
8	Afrilan	.546	.814	.803	.807
9	HFT	.477	.786	.793	.787
		Unseen Questions 2-way			
1	IWM-DKM	**.550**	**.813**	**.818**	**.815**
2	SDPA	.535	.806	.804	.805
3	WSE Research	.533	.806	.811	.808
4	RETUYT-INCO	.490	.787	.791	.789
5	HFT	.482	.786	.793	.788
6	ASLAN	.457	.780	.789	.779
7	Afrilan	.437	.766	.757	.761
8	AMATI	.289	.704	.714	.708
		Unseen Answers 3-way			
1	IWM-DKM	**.796**	**.781**	**.782**	**.780**
2	WSE Research	.790	.773	.774	.773
3	SDPA	.776	.775	.763	.766
4	ASLAN	.757	.758	.743	.746
5	AMATI	.749	.760	.736	.739
6	RETUYT-INCO	.729	.728	.728	.728
7	Diffuser	.698	.712	.703	.705
8	Afrilan	.647	.659	.661	.658
		Unseen Questions 3-way			
1	IWM-DKM	**.681**	**.680**	**.664**	**.669**
2	WSE Research	.672	.670	.663	.665
3	SDPA	.644	.634	.633	.633
4	ASLAN	.579	.653	.587	.593
5	Afrilan	.523	.607	.608	.607
6	AMATI	.394	.525	.525	.520

Table 4: The best result achieved by each team across the four evaluation tracks.

performance of fine-tuned decoder-based LLMs in these settings. The sequential combination of symbolic formulations used to prompt an LLM, as employed by *AMATI*, further trails these approaches.

In the Unseen Questions setting, however, the ranking partially differed. While *IWM-DKM* remained the strongest, the advantage of fine-tuning diminished, and approaches emphasising cross-question generalisation became more competitive: *SDPA* and *WSE Research* remained strong, but *RETUYT-INCO* and especially *HFT* closed the gap despite relying on prompt-based strategies. This highlights that without question-specific training signals, simpler, more generic approaches can likely perform comparably to complex systems. Notably, *AMATI* showed a pronounced drop in this setting. Across all tracks, *IWM-DKM*'s consistent first-place performance suggests that combining structured reasoning, task-specific fine-tuning, and ensembling provided the most robust overall solution.

7 Synthesis

Across submissions, a clear structure emerges that can be interpreted along three main axes, with individual teams occupying different positions in this space.

First, several teams converged on model-centric, LLM-driven approaches that are heavily structured and optimised. This is most evident for *IWM-DKM*, *WSE Research*, and *SDPA*. All three relied on fine-tuned LLMs as the core component, but differed in how much additional structure they imposed: *IWM-DKM* explicitly injected reasoning structure via checklist thinking and rubric reframing, *WSE Research* focused on combining multiple model

variants and retrieval signals in a unified pipeline, and *SDPA* emphasised parameter-efficient adaptation and showed that carefully fine-tuned models can outperform more complex ensembles, particularly in generalisation scenarios. Despite these differences, the shared insight is that strong performance emerged when the LLM was not used "as-is", but is systematically aligned with rubric structure, either through input design, fine-tuning, or both.

Second, a number of teams moved towards hybrid pipelines that combine complementary paradigms. This is most clearly represented by *ASLAN* and, to some extent, *WSE Research* and *SDPA*. Here, the task was decomposed into multiple views, e.g. similarity-based retrieval, classical classification, or LLM-based scoring, which are then aggregated via voting or stacking. *ASLAN* is the most explicit example of this philosophy, showing that different paradigms dominate under different conditions (e.g., similarity-based methods for seen questions vs. LLMs for unseen ones). This reinforces a broader pattern also visible in *WSE Research*: performance gains were often due to complementarity between models, rather than improvements within a single model class.

Third, several approaches focused less on the model itself and more on how the task is interfaced with or augmented around the model. *RETUYT-INCO* is the clearest example here, shifting the optimisation problem from model parameters to prompt generation, effectively learning task-specific scoring functions via meta-prompting. *AMATI* explored a different direction by injecting symbolic structure into the LLM via ILP-derived rules, trading some flexibility for interpretability and controllability. Finally, *HFT* represents the extreme end of this axis, deliberately using minimal, zero-shot LLM setups and compensating through a hybrid human–AI process in which model predictions were selectively reviewed.

8 Discussion

The results of the shared task highlight both the potential and the current limitations of rubric-based short-answer scoring. Across all tracks, approaches that explicitly align model behaviour with rubric structure achieved the strongest performance. In particular, fine-tuned systems with structured input representations consistently outperformed prompting-based approaches when sufficient task-specific information was available. However, results on the unseen question tracks showed that this advantage did not fully transfer to more challenging generalisation scenarios. While top-performing systems remained competitive, the performance gap to simpler methods narrowed. In this context, prompt-based and zero-shot methods could provide a competitive alternative despite lower peak performance. Another key observation is the importance of hybrid approaches. Many submissions combined multiple components via ensembling, integrating different scoring paradigms, or incorporating symbolic elements. This suggests that rubric-based scoring likely benefits most from combining complementary signals rather than relying on a single modelling paradigm.

9 Conclusion

Overall, the results show that structured and task-adapted LLM systems currently achieve the strongest performance, especially when questions are seen during training. At the same time, the unseen question tracks demonstrate that generalisation is possible to a certain degree, but remains challenging. We hope that this shared task provides a foundation for future work on rubric-based scoring, especially with respect to better generalisation, more interpretable scoring processes, and practically useful systems for educational settings.

Acknowledgements

This work is based on data from the *ALICE* project, funded by the *Leibniz Association* (*K365/2020*).

Limitations

Several limitations of the shared task should be considered when interpreting the results. First, the datasets used in this task are limited in scope with respect to domain, rubric complexity and score granularity. The dataset was collected in the German K12 context from grades 7-11 in the state of Schleswig-Holstein. All items use relatively simple holistic rubrics with three performance levels, which restricts conclusions about how well current approaches scale to more complex, analytic rubrics commonly used in real-world assessment.

Second, while the inclusion of unseen question tracks allows us to study generalisation, the evaluation is still confined to a single dataset and domain distribution. As a result, it remains unclear to what extent the observed findings transfer to

other subjects, educational levels, or languages beyond the German setting considered here. Third, the shared task focuses primarily on predictive performance and does not systematically evaluate the interpretability or faithfulness of scoring decisions. Although some submitted systems explicitly aim to improve transparency, the current evaluation setup does not allow for a rigorous comparison of how well models align their decisions with rubric criteria.

Several limitations of the shared task arise from constraints in dataset construction and release. First, in some cases, the provided rubrics are underspecified for the full range of valid student answers. While this reflects authentic assessment practice to some degree, it can render the scoring task inherently ambiguous and may disadvantage models that rely on precise alignment between rubric criteria and responses. Second, for a subset of questions, relevant context materials could not be included due to legal restrictions. As a result, models were required to score answers without access to information that human raters would typically consider, which likely limited achievable performance and may have introduced additional variance across items.

Finally, while the shared task setup aims to approximate realistic assessment conditions, the combination of partial rubric specification and missing context materials means that the benchmark does not fully capture ideal scoring conditions for all questions.

Ethics Statement

With this shared task, we aimed to explore the potential of rubric-based scoring for short answers. Since short-answer scoring directly concerns the rating of students' performance, the ethicality of the technology depends heavily on the specific setting in which it is deployed. Technology for rating students' performance should match or outperform human performance when deployed for high-stakes assessments and be subject to human supervision. The EU AI Act labels AI in education as overall high stakes. However, we argue that there are also low-stakes scenarios in which rubric-based short answer scoring technologies can be deployed. These include typical low-stakes learning scenarios as provided by intelligent tutoring systems or comparable applications. Overall, significant downstream research on technology for rubric-based short-answer scoring is likely needed before the respective technology can be deployed in practical scenarios.

References

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

David Bednorz, Kristin Litteck, Daniel Sommerhoff, and Aiso Heinze. 2024. *Erfassung individueller Lerntrajektorien in einer digitalen Lernumgebung zum Ableitungsbegriff*. Universitätsbibliothek Dortmund.

Kate Rebecca Belcher, Marius De Kuthy Meurers, Kordula De Kuthy, and Detmar Meurers. 2026. IWM-DKM at BEA 2026 shared task 2: Supplementing supervised fine-tuning for rubric-based short answer scoring. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

Marie Bexte, Yuning Ding, Josef Ruppenhofer, Nils-Jonathan Schaller, Daniel Mora Melanchthon, Torsten Zesch, and Andrea Horbach. 2026. ASLAN at BEA 2026 shared task 2: Voting across scoring paradigms. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

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.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2023. Similarity-based content scoring - a more classroom-suitable alternative to instance-based scoring? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1892–1903, Toronto, Canada. Association for Computational Linguistics.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2024. Strengths and weaknesses of automated scoring of free-text student answers. *Informatik Spektrum*, pages 1–9.

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.

Leon Camus and Anna Filighera. 2020. Investigating transformers for automatic short answer grading. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 43–48. Springer.

Aubrey Condor, Zachary Pardos, and Marcia Linn. 2022. Representing scoring rubrics as graphs for automatic short answer grading. In *International Conference on Artificial Intelligence in Education*, pages 354–365. Springer.

Longwei Cong, Leon Hammerla, Sonja Hahn, Sebastian Gombert, Hendrik Drachsler, and Ulf Kroehne. 2026. Automatic short answer grading with llms: From memorization to reasoning. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, pages 75–84.

Berrit K Czinczel, Daniela Fiedler, and Ute Harms. 2025. How do species change over time? designing a hybrid teaching unit on five factors of evolution. *The American Biology Teacher*, 87(2):78–83.

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.

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 (Volume 1: Long Papers)*, pages 8577–8591.

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 *International Conference on Artificial Intelligence in Education*, pages 44–51. Springer.

Sebastian Gombert, Daniele Di Mitri, Onur Karademir, Marcus Kubsch, Hannah Kolbe, Simon Tautz, Adrian Grimm, Isabell Bohm, Knut Neumann, and Hendrik Drachsler. 2023. Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning*, 39(3):767–786.

Sebastian Gombert, Sonja Hahn, Nico Andersen, Leon Camus, Zhifan Sun, Ngoc Nhu Hao Nguyen, Fabian Zehner, Longwei Cong, Alexander Mehler, and Hendrik Drachsler. 2026a. Rubrics as semantic subspaces: A unified approach to rubric-based constructed response scoring across short answers and essays. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

Sebastian Gombert, Zhifan Sun, Fabian Zehner, Jannik Lossjew, Tobias Wyrwich, Berrit Katharina Czinczel, David Bednorz, Marcus Kubsch, Daniele Di Mitri, Knut Neumann, and 1 others. 2026b. Are rubrics all you need? towards rubric-based automatic short answer scoring via guided rubric-answer alignment. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, pages 272–282.

Jonas Gwozdz and Andreas Both. 2026. WSE research at BEA 2026 shared task 2: Multi-strategy rubric-based short answer scoring for german: Multi-strategy 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.

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring. https://kaggle.com/competitions/asap-aes. Kaggle.

Marcus Kubsch, Berrit Czinczel, Jannik Lossjew, Tobias Wyrwich, David Bednorz, Sascha Bernholt, Daniela Fiedler, Sebastian Strauß, Ulrike Cress, Hendrik Drachsler, and 1 others. 2022. Toward learning progression analytics—developing learning environments for the automated analysis of learning using evidence centered design. In *Frontiers in education*, volume 7, page 981910. Frontiers Media SA.

Zhaohui Li, Susan Lloyd, Matthew Beckman, and Rebecca J Passonneau. 2023. Answer-state recurrent relational network (AsRRN) for constructed response assessment and feedback grouping. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3879–3891.

Zhexiong Liu and Jing Zhang. 2026. SDPA at BEA 2026 shared task 2: Efficient LLM fine-tuning for rubric-based short answer scoring. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

Jannik Lossjew and Sascha Bernholt. 2025. Digital support custom-made or in moderation–design and evaluation of a digitally supported teaching unit on reaction kinetics and chemical equilibrium. *CHEMKON*, 32(4):122–131.

OECD. 2023. *PISA 2022 Assessment and Analytical Framework*. PISA. OECD Publishing.

OECD. 2024. *PISA 2022 Technical Report*. PISA. OECD Publishing.

Ulrike Padó. 2026. HFT at BEA 2026 shared task 2: Blunt-edge models for hybrid grading. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

Ulrike Padó, Yunus Eryilmaz, and Larissa Kirschner. 2023. Short-answer grading for german: Addressing the challenges. *International Journal of Artificial Intelligence in Education*, 34(4):1321–1352.

Ernesto Panadero and Anders Jonsson. 2013. The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational research review*, 9:129–144.

Y Malini Reddy and Heidi Andrade. 2010. A review of rubric use in higher education. *Assessment & evaluation in higher education*, 35(4):435–448.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Ignacio Sastre, Ignacio Remersaro, Facundo Díaz, Nicolás de Horta, Luis Chiruzzo, Aiala Rosá, and Santiago Góngora. 2026. RETUYT-INCO at BEA 2026 shared task 2: Meta-prompting in rubric-based scoring for german. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

Shashank Sonkar, Kangqi Ni, Lesa Tran Lu, Kristi Kincaid, John S Hutchinson, and Richard G Baraniuk. 2024. Automated long answer grading with ricechem dataset. In *International Conference on Artificial Intelligence in Education*, pages 163–176. Springer.

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.

Yuchen Wei, Dennis Pearl, Matthew Beckman, and Rebecca J Passonneau. 2025. Concept-based rubrics improve LLM formative assessment and data synthesis. *arXiv preprint arXiv:2504.03877*.

Alistair Willis and Aisling Third. 2026. AMATI at BEA 2026 shared task 2: Automatic short answer grading with inductive logic programming and a large language model. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

Tobias Wyrwich, Marcus Kubsch, Hendrik Drachsler, and Knut Neumann. 2025. Tracking students' progression in developing understanding of energy using ai technologies. *Physical Review Physics Education Research*, 21(1).

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7).

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

A Appendix

Rank	Team Name	↑ Quadratic Weighted Kappa	Weighted Precision	Weighted Recall	Weighted F1-Score
1	IWM-DKM	.726	.887	.887	.887
2	WSE Research	.717	.882	.883	.883
3	WSE Research	.717	.882	.883	.883
4	IWM-DKM	.710	.880	.881	.881
5	WSE Research	.710	.880	.880	.880
6	WSE Research	.702	.878	.879	.878
7	IWM-DKM	.700	.876	.877	.876
8	IWM-DKM	.698	.875	.877	.876
9	WSE Research	.697	.874	.875	.874
10	IWM-DKM	.684	.869	.870	.869
11	SDPA	.682	.869	.866	.867
12	SDPA	.679	.867	.866	.866
13	RETUYT-INCO	.674	.865	.861	.863
14	RETUYT-INCO	.673	.866	.861	.862
15	RETUYT-INCO	.672	.864	.862	.863
16	RETUYT-INCO	.667	.862	.862	.862
17	ASLAN	.663	.861	.864	.862
18	SDPA	.662	.860	.862	.860
19	SDPA	.656	.857	.856	.857
20	ASLAN	.654	.860	.863	.859
21	ASLAN	.654	.857	.860	.858
22	RETUYT-INCO	.654	.858	.851	.853
23	SDPA	.652	.856	.858	.856
24	ASLAN	.648	.857	.860	.856
25	AMATI	.644	.857	.843	.847
26	AMATI	.644	.857	.843	.847
27	ASLAN	.637	.850	.853	.851
28	Diffuser	.615	.840	.842	.841
29	Afrilan	.546	.814	.803	.807
30	AMATI	.489	.801	.807	.795
31	AMATI	.485	.802	.807	.794
32	AMATI	.485	.802	.807	.794
33	HFT	.477	.786	.793	.787
34	HFT	.465	.781	.789	.783
35	HFT	.435	.766	.770	.768
36	HFT	.424	.761	.766	.763

Table 5: The best five runs per team for the Unseen Answers 2-way track.

Rank	Team Name	↑ Quadratic Weighted Kappa	Weighted Precision	Weighted Recall	Weighted F1-Score
1	IWM-DKM	.550	.813	.818	.815
2	SDPA	.535	.806	.804	.805
3	WSE Research	.533	.806	.811	.808
4	SDPA	.531	.806	.811	.807
5	IWM-DKM	.526	.802	.797	.799
6	WSE Research	.525	.801	.804	.803
7	WSE Research	.510	.799	.806	.800
8	WSE Research	.503	.792	.792	.792
9	WSE Research	.503	.792	.792	.792
10	IWM-DKM	.501	.796	.804	.797
11	RETUYT-INCO	.49	.787	.791	.789
12	HFT	.482	.786	.793	.788
13	HFT	.477	.784	.791	.785
14	HFT	.467	.780	.787	.782
15	SDPA	.461	.776	.781	.778
16	ASLAN	.457	.780	.789	.779
17	ASLAN	.454	.778	.787	.778
18	HFT	.452	.772	.763	.767
19	SDPA	.452	.771	.772	.772
20	ASLAN	.452	.772	.779	.774
21	HFT	.447	.770	.760	.764
22	Afrilan	.437	.766	.757	.761
23	SDPA	.436	.767	.775	.769
24	RETUYT-INCO	.432	.764	.770	.766
25	ASLAN	.417	.770	.780	.765
26	ASLAN	.388	.747	.757	.75
27	RETUYT-INCO	.341	.746	.760	.736
28	AMATI	.289	.704	.714	.708

Table 6: The best five runs per team for the Unseen Questions 2-way track.

Rank	Team Name	↑ Quadratic Weighted Kappa	Weighted Precision	Weighted Recall	Weighted F1-Score
1	IWM-DKM	.796	.781	.782	.780
2	IWM-DKM	.790	.772	.77	.771
3	WSE Research	.790	.773	.774	.773
4	WSE Research	.790	.769	.768	.768
5	WSE Research	.788	.773	.773	.773
6	WSE Research	.784	.765	.765	.765
7	WSE Research	.781	.765	.765	.765
8	IWM-DKM	.781	.779	.767	.77
9	IWM-DKM	.78	.775	.768	.77
10	IWM-DKM	.779	.771	.763	.765
11	SDPA	.776	.775	.763	.766
12	SDPA	.758	.746	.749	.744
13	SDPA	.757	.767	.750	.753
14	ASLAN	.757	.758	.743	.746
15	SDPA	.757	.770	.752	.755
16	SDPA	.756	.760	.755	.756
17	ASLAN	.754	.757	.741	.744
18	AMATI	.749	.760	.736	.739
19	ASLAN	.747	.750	.737	.740
20	ASLAN	.744	.744	.728	.731
21	ASLAN	.739	.740	.725	.728
22	RETUYT-INCO	.729	.728	.728	.728
23	AMATI	.721	.755	.711	.715
24	Diffuser	.698	.712	.703	.705
25	RETUYT-INCO	.695	.701	.702	.702
26	Afrilan	.647	.659	.661	.658
27	AMATI	.614	.662	.655	.650
28	AMATI	.567	.645	.633	.629
29	AMATI	.491	.596	.596	.595

Table 7: The best five runs per team for the Unseen Answers 3-way track.

Rank	Team Name	↑ Quadratic Weighted Kappa	Weighted Precision	Weighted Recall	Weighted F1-Score
1	IWM-DKM	.681	.680	.664	.669
2	WSE Research	.672	.670	.663	.665
3	WSE Research	.653	.664	.662	.663
4	SDPA	.644	.634	.633	.633
5	SDPA	.636	.650	.627	.633
6	IWM-DKM	.635	.657	.640	.644
7	SDPA	.629	.644	.625	.630
8	SDPA	.627	.646	.617	.624
9	WSE Research	.625	.650	.650	.650
10	WSE Research	.625	.655	.647	.648
11	WSE Research	.621	.649	.650	.649
12	SDPA	.621	.630	.628	.628
13	IWM-DKM	.591	.727	.601	.600
14	ASLAN	.579	.653	.587	.593
15	ASLAN	.542	.701	.576	.570
16	IWM-DKM	.541	.738	.576	.563
17	ASLAN	.539	.676	.571	.567
18	ASLAN	.525	.643	.566	.567
19	Afrilan	.523	.607	.608	.607
20	ASLAN	.517	.609	.562	.568
21	AMATI	.394	.525	.525	.520

Table 8: The best five runs per team for the Unseen Questions 3-way track.