ALICE: A Large-Scale German Benchmark for Rubric-Based Multi-Dimensional Automatic Short Answer Scoring

Anonymous ACL submission

Abstract

Automatic Short Answer Scoring (ASAS) is central to NLP for Education. However, openly available benchmarks remain scarce, and existing datasets largely assess how well students answer a question directly rather than how well they master underlying **concepts** (knowledge elements) such as *thermal energy* or **epistemic activities** (skill) such as reasoning or claim. To address this gap, we introduce ALICE, a large-scale German ASAS dataset comprising three subtasks: (i) learning performance (ALICE-LP), (ii) knowledge elements (ALICE-KE), and (iii) skills (ALICE-SK). We further frame ASAS as a rubric-ranking task and benchmark with a range of language models, from small MLMs to several lightweight LLMs, under various input configurations. Our experiments show that lightweight LLMs used as encoders are particularly effective for rubric-based ASAS. We also investigate what combinations of context information (rubrics, prompts, sample solutions) are beneficial for ASAS.

1 Introduction

Automatic short answer scoring (ASAS) is a fundamental task in NLP for Education. Extensive research has focused on both algorithmic approaches (Bexte et al., 2022; Li et al., 2023; Zehner et al., 2025; Wang et al., 2019; Bexte et al., 2024) and the development of benchmark datasets (Dzikovska et al., 2013; Filighera et al., 2022). In real-world educational settings, short-answer questions are designed to probe students' understanding of concepts taught during instruction. By constructing answers of their own, students can actively reflect on the taught knowledge and foster critical thinking (Bai and Stede, 2023). However, manually grading students' answers is tedious and expensive. Thus, developing benchmarks and models that automate this process can significantly alleviate the workload of the teachers and ensure timely and individualised feedback.

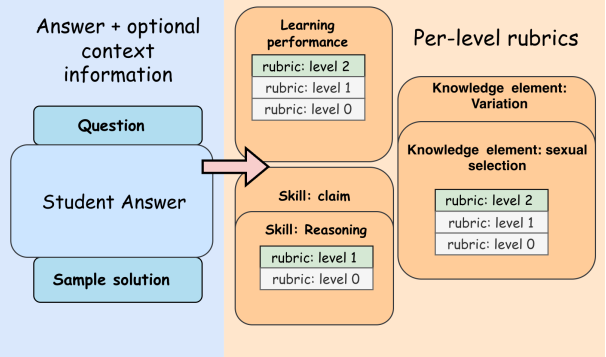

Figure 1: Structure of ALICE

Despite substantial progress, existing ASAS benchmarks and models remain misaligned with how human teachers evaluate student responses. Most datasets adopt a solution-based formulation, in which answers are scored by their similarity to a reference solution or by selecting from a fixed set of labels (Sung et al., 2019; Filighera et al., 2022; Camus and Filighera, 2020; Ormerod, 2022; Kumar et al., 2019). This design implicitly assumes that scoring criteria are uniform across questions. In practice, however, questions may have varying performance levels depending on the pedagogical need, and teachers rely on question-specific rubrics (Brookhart, 2018; Panadero and Jonsson, 2013; Krebs et al., 2022) for scoring. As a result, standard classification-based approaches struggle to generalise to unseen questions.

In terms of the scoring objectives, knowledge elements (KE) and learning skills (SK) are critical in STEM contexts, where the ultimate goal of teaching is to master scientific concepts and apply epistemic activities to solve problems (Stanja et al., 2023; Gombert et al., 2023; Reynders et al., 2020). Yet, all existing ASAS benchmarks only evaluate how students directly address the questions.

To address this gap, we introduce rubric-ranking as a unifying task formulation for ASAS, in which student responses are scored by directly aligning them with question-specific rubric items rather than predicting fixed labels. We also present ALICE (Figure 1), a large-scale, German-language ASAS benchmark. ALICE is designed to reflect the multi-dimensional nature of real-world educational assessment. It consists of three sub-tasks: ALICE-LP (learning performance), which measures how well students answer the question; ALICE-KE (knowledge element), where understanding of target concepts is scored; and ALICE-SK (skills), where students' problem-solving skills, as demonstrated in the answers, are scored. Crucially, every instance in ALICE is accompanied by the full grading context: the *question prompt*, *per-level rubric*, and a *sample solution*—closely aligned with how human teachers evaluate responses. This design allows us to test the extent to which rubric-based ASAS models rely on these contextual components to score responses effectively. (See Appendix A for the complete example of the dataset.)

For the base model, we adopt large language models (LLMs) as encoders, motivated by several factors. First, LLMs exhibit strong reasoning capabilities and massive subject-matter knowledge (Tay et al., 2023; Han et al., 2025; Wei et al., 2022), which are crucial for ASAS in STEM subjects. Second, their ability to generalise makes them well-suited for handling unseen questions, a common challenge in ASAS. Finally, our dataset design involves *multi-sequence inputs*: question prompts, answers, rubrics, and sample solutions. Recent studies such as Ruan et al. (2024) have shown that LLMs can flexibly adapt to multiple input sequences.

The contribution of our work can be summarised as follows:

- We introduce ALICE, a challenging German ASAS benchmark comprising three complementary subtasks. The dataset provides rich, human-facing context—including prompts, per-level rubrics, and sample solutions—that closely reflects real-world scoring practices and pedagogical objectives in STEM education.
- We approach ASAS as rubric-ranking and demonstrate consistent advantages of rubric-based scoring over solution-based approaches across datasets and splits. It achieves stronger performance and scales effectively across questions with varying difficulty levels.
- We conduct extensive experiments with masked language models (MLM) and LLM under multi-sequence input settings to systematically assess which contextual components are essential for ASAS. This provides valuable guidelines for the design of future ASAS datasets and models.

2 Background

Short Answer Scoring Datasets Despite its importance in NLP for Education, publicly available non-English ASAS benchmarks remain scarce, and existing benchmarks predominantly assess how well student responses address the question, rather than their mastery of underlying concepts or skills. SciEntsBank (Dzikovska et al., 2013) and ASAP-SAS are the most widely used datasets in prior work (Riordan et al., 2017; Bexte et al., 2022; Wang et al., 2019; Ormerod, 2022; Kumar et al., 2019). For multilingual evaluation, there exist the German SAF (Filighera et al., 2022) and the Portuguese PT_ASAG_2018 (Galhardi et al., 2018). In Sonkar et al. (2024), the author proposes *Long Answer Scoring* where the student's answer consists of a paragraph, in contrast to short answers, which are usually 2-3 sentences long. However, due to privacy constraints, a large amount of work in this field evaluates their scoring models on proprietary datasets (Chang et al., 2024; Sung et al., 2019; OECD-PISA, 2023).

Most benchmarks in ASAS mentioned above adopt a solution-based design, where the student's answer is scored by comparing it with the reference answer. While straightforward to implement, solution-based scoring only measures the similarity to the highest-level answer and does not provide signals regarding incorrect or partially correct answers. Such responses often reveal valuable insights into students' learning progress (Sadler, 1989; Fisher and Lipson, 1986). To our best knowledge, ASAP-SAS and Sonkar et al. (2024) are the only rubric-based ASAS benchmarks. The former assumes an **exclusive** relation between the rubrics. That is, one answer can only satisfy one rubric. In the latter one, multiple rubrics can be satisfied and the final score of an answer is obtained by summing the matched rubrics. Our work follows the design of the ASAP-SAS. A design that lies somewhere between the solution-based and the rubric-based

Dataset Name	Size	Context Information	Language Coverage	Public?
ASAP-SAS	22K	per-level rubrics	English	✓
ScienceBank (Dzikovska et al., 2013)	10K	solution, question prompt	English	✓
PT_ASAG_2018 (Galhardi et al., 2018)	13k	solution, question prompt	Portuguese	✓
Sung et al. (2019)	76k	solution, question prompt	English	×
SAF (Filighera et al., 2022)	4.5k	solution, question	German, English	✓
ISTUDIO Li et al. (2023)	6.5k	question context, question prompt, per-level reference answer	English	✓
Chang et al. (2024)	76k	question prompt	Finnish	×
(Sonkar et al., 2024)	1265	per-level rubric	English	✓
ALICE (ours)	16k	question prompt, solution, per-level rubrics	German	✓

Table 1: Comparison of ALICE with existing ASAS benchmarks. Note that we frame all other information other than student's answer as *context information*.

ones is ISTUDIO [1]. It contains a list of representative answers for each performance level to provide per-level signals.

The ALICE benchmark provides per-level rubrics for a three-dimensional evaluation of student responses, enabling a multi-faceted assessment that jointly captures correctness, conceptual understanding, and the reflected cognitive skills.

LLMs beyond generation There is growing interest in using LLMs for embedding and classification. This line of work aims to compensate for the limitations of MLMs (e.g. Bert, Roberta, etc.), particularly their limited domain knowledge and sample efficiency (Luo et al., 2024; BehnamGhader et al., 2024; Liu et al., 2024). Several studies further modify LLM architectures to overcome their autoregressive nature. For example, converting attention from causal to bidirectional (BehnamGhader et al., 2024; Lin et al., 2025) or inserting transformer adapters on top of hidden states (Lee et al., 2025). Liu et al. (2024) further explored the combination of hidden states from multiple LLM layers and their integration with representations from MLMs, yielding improved performance on various sequence classification tasks.

Most prior work evaluating LLMs as encoders relies on MTEB (Muennighoff et al., 2023) as the evaluation benchmark, which consists of clustering, re-ranking, and classification tasks. Although it includes sequence-pair classification (e.g., duplicate question detection), these tasks are treated as standard embedding evaluations: the two sequences are encoded independently, and predictions are derived from cosine or dot product similarity. This is fundamentally mismatched with rubric-based ASAS, where scoring depends on joint reasoning over answers, rubrics and optionally more context information.

Ruan et al. (2024) systematically examined architectures and input structures for LLM-based classification on Edit Intent Classification (EIC), a sequence-pair task. It shows that LLMs are competitive multi-sequence classifiers. Compared to EIC, rubric-based ASAS involves more diverse and independent input sequences and longer context windows.

Given the above-mentioned properties LLM encoders are particularly well-suited to our setting. We therefore conduct comprehensive experiments on four lightweight LLMs across multiple input formats.

3 The ALICE Dataset

We introduce ALICE, which stands for [blinded], a large-scale German ASAS dataset that augments prior learning-performance annotations with fine-grained scoring of knowledge elements and skills. Each question contains three types of rubrics:

- **Learning Performance (LP)**: the degree to which a student's response directly satisfies the question prompt. There are three performance levels: correct, partially correct, and incorrect.
- **Knowledge elements (KE)**: the extent to which a student's response correctly employs the targeted domain concepts specified for the question. Typically, there are four levels: *No Use*; *Use without content*; *Non-targeted use* and *Targeted use*. The rubric of *Use without content* are unavailable for some questions.
- **Skills (SK)**: the presence of specified cognitive or reasoning behaviours in a student's response. Typically, the levels are: *Present*

[1] https://www.datacommons.psu.edu/commonswizard/MetadataDisplay.aspx?Dataset=6392

and *Not present*. For some questions, *Partially present* is possible.

These dimensions were chosen to reflect complementary aspects of assessment commonly used in STEM education: task completion (LP), conceptual understanding (KE), and epistemic activity (SK).

Split	Task	# of Questions	# of Answers	# of Items
Train	LP	90	10783	10783
	KE	78	9734	24012
	SK	80	9978	11644
TEST-UQ	LP	22	3096	3096
	KE	17	2598	6505
	SK	17	2595	3105
TEST-UA	LP	90	2695	2695
	KE	78	2442	5874
	SK	80	2494	2915

Table 2: Dataset Statistics

After annotation, we filter out questions without sample solutions, e.g. open-ended questions, to normalise the dataset structure. The result is a corpus of 16,574 student answers. To ensure robust generalisation, we further split the test set into TEST-UA (unseen answers) and TEST-UQ (unseen questions). The detailed statistics are in Table 2.

An answer may contain **multiple** KEs and SKs, and we do not predict the scores for multiple KEs/SKs jointly during training and testing. Hence, we extend the answer by the number of KEs/SKs to answer-KE/SK pair. Each pair is treated as an independent instance during training and testing. This is indicated by the *number of items* in Table 2.

We observe strong positive correlations between the scores of Learning Performance and both other dimensions: LP-KE (r=0.785, p<0.001) and LP-Skills (r=0.631, p<0.001). The stronger correlation with knowledge elements indicates a closer empirical association between LP and KE than between LP and SK, though all dimensions capture distinct aspects of student responses. These results validate our multidimensional framework by demonstrating that while the dimensions are related, they capture distinct aspects of student understanding. For instance, a student who fails to answer the question correctly might still demonstrate sufficient epistemic abilities and mastery of concepts of interest. This provides more fine-grained information on the learning progress of the students. We report the data collection process, annotation protocol, and dataset examples in Appendix C.

4 Approach

4.1 Task Formulation

Each instance in the ALICE dataset is represented as a tuple $\langle q, a, s, R\rangle$, where q denotes the question prompt, a the student response, s a sample solution, and R the set of rubric items associated with the instance. The rubric set R is question-specific and varies in size depending on the subtask.

The structure of R differs across subtasks. For **ALICE-LP**, the rubric set

$$R_q = \{r_0^{(q)}, r_1^{(q)}, r_2^{(q)}\}$$

corresponds to the ordered performance levels defined for question q, where each rubric item describes the qualitative criteria for one score level.

For **ALICE-KE** and **ALICE-SK**, each question q is annotated with a set of targets

$$T_q = \{t_1, \ldots, t_m\},$$

where each target t_j corresponds to a knowledge element or a skill. Each target t_j is associated with its own rubric set $R_{t_j}^{(q)}$, which defines the scoring criteria for that specific concept or skill. In practice, each answer–target pair induces a small, independent rubric set that is ranked separately.

Across all subtasks, the model's objective is to determine which rubric items in R are satisfied by a student response a. We formulate this as a *rubric-ranking* task, where rubric items are ranked conditioned on the combined input $\langle q, a, s, r\rangle$. Ranking enables direct comparison between rubric levels without assuming mutually exclusive labels or uniformly spaced score categories, which is particularly important given the ordinal and subsumptive structure of educational rubrics.

Input Format. For our proposed rubric-ranking approach, only *a* and *r* are the essential input documents to the encoder, but we wish to investigate further how incorporating extra context information affects model performance. We perform extra experiments where the question prompt (+q), sample solution (+s) and their combination (+qs) are input to the model[2].

4.2 Modelling

We employ a language model (LM) as the encoder Φ_θ to encode the aforementioned input sequences.

[2]While many permutations of the input sequence are possible, in our experiment we uniformly attach the context information to the front of the model. That is: *qar*, *sar*, *qsar*

For MLMs (Devlin et al., 2019), we separate segments using the [SEP] token. For LLM-based ones, we adopt the best-performing *structured* format proposed by (Ruan et al., 2024); implementation details are provided in the Appendix B.

The combined input sequence is fed into the encoder Φ_θ, which outputs the contextualised token embeddings. For BERT-base models, we extract the representation of the [CLS] token; for LLM-based models, we use the final [EOS] token representations. This sequence embedding is then passed through a feedforward layer to produce a scalar alignment score:

$$s_{i,j} = f_\theta(q_i, a_i, s_i, r_j),$$

where $s_{i,j} \in \mathbb{R}$ denotes how well rubric item r_j matches the student answer a_i, optionally conditioned on q_i and s_i.

It is also possible to encode the answer and rubric separately and compute their alignment score via cosine similarity or dot product, following the standard approach in sentence embeddings and ranking task (Reimers and Gurevych, 2019; Muennighoff et al., 2023). We won't adopt this approach because the joint encoding of multiple sequences in a single sequence enables richer cross-attention and interaction among segments. The joint modelling strategy was also shown to outperform separate encoding in pairwise classification tasks (Ruan et al., 2024). In addition, since there are only a few rubric candidates to rank for each answer, joint encoding won't lead to excessive computation and latency, in contrast to retrieving tens or hundreds of potential documents.

Objective Functions. The model is trained to predict a scalar alignment score $s_{i,j} \in [0, 1]$ indicating the degree of match between a_i and r_j, optionally conditioned on q_i and s_i. We experiment with two training objectives:

(1) Contrastive Loss (CL). For each student–rubric pair (a_i, r_j) with label $y_{i,j} \in \{0, 1\}$ and model output $s_{i,j} \in (0, 1)$, the contrastive loss for this item is:

$$\ell_{\text{contrast}}^{(i,j)} = \begin{cases} (1 - s_{i,j})^2 & \text{if } y_{i,j} = 1 \\ (s_{i,j})^2 & \text{if } y_{i,j} = 0 \end{cases} \quad (1)$$

(2) Softmax Cross-Entropy Loss. Alternatively, we formulate rubric-ranking as a multi-class classification problem within each rubric set. Let $P_\theta(j \mid i)$ denote the probability assigned to rubric item r_j for student answer a_i, obtained by applying a softmax over the alignment scores for all rubric items associated with the same instance:

$$P_\theta(j \mid i) = \frac{\exp(s_{i,j})}{\sum_{k=1}^{|R|} \exp(s_{i,k})}. \quad (2)$$

The softmax cross-entropy loss for the correct rubric item r_j is then defined as:

$$\ell_{\text{SCE}}^{(i,j)} = -\log P_\theta(j \mid i). \quad (3)$$

The softmax normalisation is computed over the rubric set of a single question or target, which typically contains only a small number of items.

Solution-based Baseline for ALICE We adopt the standard modelling strategy used in prior ASAS work (Sung et al., 2019; Camus and Filighera, 2020). Specifically, the student answer and the corresponding sample solution are concatenated and provided as input to a language model with a classification head. We refer to this setup as the **solution-based** approach, in contrast to our proposed **rubric-based** methods.

This baseline is applied to all ALICE subtasks. For ALICE-KE and ALICE-SK, the number of performance levels may vary across questions, with up to four levels defined for ALICE-KE. When intermediate levels such as *use without content* are absent, we preserve the original level-to-index mapping (*no use* $\rightarrow 0, \ldots,$ *targeted use* $\rightarrow 3$) and employ a classification head with num_labels$= 4$ (see § 3 and Appendix A).

Prompting Closed-source Models We prompt closed-source large language models to perform rubric-based scoring by providing the same input components used for encoder-based models, including the student response and the corresponding rubric items, optionally augmented with the question prompt and sample solution. The models are instructed to select the most appropriate rubric level for each instance. We report the detailed setup and the results in Appendix E.

5 Main Experiments

5.1 Experiment Setups

We benchmark the proposed rubric-ranking approach across all ALICE subtasks under both unseen-answer (UA) and unseen-question (UQ) settings. Experimental details, including model

configurations and hyperparameters, are provided in Appendix D. We report F1 and quadratic weighted kappa (QWK), with particular emphasis on UQ performance to assess generalisation beyond the seen ones. We also report the metric differences between TEST-UA and TEST-UQ to better illustrate their generalisation ability.

5.2 Results and Discussion

RQ1: Which loss function is better suited for the rubric-ranking? Table 3 reports the main results for the rubric-based experiments. Across models and input formats, SCE consistently outperforms CL. The discrepancy in performance is starker when it comes to TEST-UQ, indicating superior generalisation ability with SCE.

The key limitation with CL is that it does not force the model to discriminate between *levels* of rubrics. It treats rubric satisfaction as independent binary decisions and therefore fail to enforce discrimination between closely related levels.

In ALICE-LP, rubrics can be subsumptive rather than exclusive. For instance, level-2 rubrics typically take the form *the student does X and Y*, whereas level-1 rubrics are: *the student X or Y*. An answer that satisfies the first condition also satisfies the weaker latter one. Thus, without explicitly distinguishing between different levels of rubrics, the model can easily confuse these labels.

Given the superior performance of SCE, the remaining subtasks are benchmarked without CL. CL is also considerably slower: each training item expands into multiple rubric–answer pairs, and the loss is computed separately for each pair. With the scale of ALICE-KE and ALICE-SK, evaluating all input formats would be prohibitively expensive.

RQ2: What is the effect of additional input information (question, sample solution) on the results? We report the benchmarking results in Table 4 and summarise the average performance across different input formats. Across all models and subtasks, while no single input format is universally optimal, we observe a consistent positive effect of adding contextual information beyond the basic answer-only format (ar). Incorporating more context information yields clear and stable improvements in most settings. Although including the full contextual information (+qs) does not guarantee the best performance for every individual model, the best **global** results across all experiments are consistently achieved with this input format when paired with the strongest model in our benchmark, *Llama-3.2-3B-Instruct*. This suggests that larger, instruction-fine-tuned models are better able to integrate and reason over multiple input sequences, effectively exploiting the richer context.

With respect to generalisation, we observe that the impact of additional context is substantially larger on TEST-UQ than on TEST-UA. While the performance gain on unseen answers is typically modest (around 1 F1 point), the gains on unseen questions are markedly larger, indicating that contextual information can mitigate question-specific distribution shifts.

RQ3: How do models perform differently on different subtasks? By comparing Table 3 and Table 4, we can see that despite their novelty, the performance on ALICE-KE and ALICE-LP are consistently better than ALICE-LP, indicated by higher scores and lower performance differences between UA and UQ. Part of the reason is that ALICE-SK is mostly a binary classification task, with a limited number of three-level rubrics. Still, most rubrics in ALICE-KE contain four levels. The reason is that various answers might share the same knowledge elements and skills. This leads to better generalisation to unseen questions. The rubrics for knowledge elements specifically focus on the use of scientific terms, whereas the rubrics for SK identify the presence of skills such as *describing the data*. As a result, the rubrics in ALICE-KE and ALICE-SK have similar formulations across questions, while the formulation of LP rubrics is highly question-dependent. This is indicated by Figure 2, where LP rubrics are less similar to each other than KE and SK.

Finally, the effect of input formats differs substantially across subtasks. For the ALICE-LP, providing the full context (+qs) does not consistently lead to optimal performance, and simpler formats such as +q or +s can be competitive or even preferable. In contrast, for subtasks ALICE-KE and ALICE-SK, the full-context input yields more reliable strong performance across various models. When we look away from the +qs settings and look at the contribution of the question prompt and the sample solution individually, we can see that

RQ4: What is the performance difference between the rubric-based and solution-based approach? We compare rubric-based scoring with a standard solution-based baseline across all ALICE

Base Model	Input Format	SCE			CL		
		UA (F1/QWK)	UQ (F1/QWK)	Δ(UA − UQ)	UA (F1/QWK)	UQ (F1/QWK)	Δ(UA − UQ)
XLM-RoBERTa-Long	*ar*	70.3 / 68.9	56.0 / 46.7	14.3 / 22.2	69.2 / 69.0	56.6 / 44.0	12.6 / 25.0
	+q	70.8 / 70.4	59.0 / 53.5	11.8 / 16.9	70.0 / 70.6	51.4 / 38.2	18.6 / 32.4
	+s	69.8 / 69.8	55.1 / 53.7	14.7 / 14.1	63.4 / 55.4	47.4 / 28.2	16.0 / 26.2
	+qs	70.2 / 70.7	53.8 / 51.2	16.4 / 19.5	48.2 / 27.8	34.2 / 3.7	14.0 / 24.1
Llama-3.2-1B	*ar*	72.9 / 71.8	61.3 / 54.7	11.6 / 17.1	68.6 / 69.0	58.2 / 49.2	10.4 / 19.8
	+q	72.5 / 72.3	61.5 / 53.9	11.0 / 18.4	67.9 / 69.3	57.6 / 49.6	10.3 / 19.7
	+s	72.8 / 72.8	61.9 / 56.1	10.9 / 10.7	66.9 / 66.1	56.1 / 52.5	10.8 / 13.6
	+qs	71.8 / 71.3	59.1 / 54.1	12.7 / 17.2	65.8 / 66.4	52.8 / 35.1	13.0 / 31.3
Llama-3.2-1B-Instruct	*ar*	71.9 / 71.0	62.0 / 54.7	9.9 / 16.3	69.2 / 70.3	58.2 / 50.0	11.0 / 20.3
	+q	72.5 / 72.2	62.4 / 56.4	10.1 / 15.8	69.3 / 70.0	54.8 / 47.0	14.5 / 23.0
	+s	71.7 / 72.5	59.5 / 59.7	12.2 / 12.8	72.1 / 73.2	56.4 / 52.2	15.7 / 21.0
	+qs	72.7 / 72.4	61.1 / 54.8	11.6 / 17.6	70.8 / 72.0	57.2 / 52.5	13.6 / 19.5
Llama-3.2-3B	*ar*	73.9 / 73.9	63.4 / 58.5	10.5 / 15.4	72.4 / 72.8	61.0 / 56.0	11.4 / 16.8
	+q	74.0 / 75.2	67.3 / 64.8	6.7 / 10.4	72.7 / 74.3	62.9 / 60.6	**9.8** / 13.7
	+s	74.2 / 75.7	63.6 / 58.9	10.3 / 16.8	73.9 / 75.7	62.6 / 59.0	11.3 / 16.7
	+qs	73.7 / 74.4	65.8 / 64.2	7.9 / 10.2	71.8 / 72.0	59.1 / 57.0	12.7 / 15.0
Llama-3.2-3B-Instruct	*ar*	73.9 / 73.8	64.0 / 60.2	9.9 / 13.6	72.9 / 73.9	59.1 / 52.9	13.8 / 21.0
	+q	74.8 / 75.5	66.0 / 62.2	8.8 / 13.3	74.0 / 75.3	60.7 / 57.1	13.3 / 18.2
	+s	73.6 / 75.3	67.2 / **67.9**	**6.4** / **7.4**	75.0 / 77.4	**64.2** / **64.2**	10.8 / **13.2**
	+qs	**75.8** / **76.5**	**67.6** / 65.5	8.2 / 11.0	**75.8** / **78.1**	62.0 / 59.1	13.8 / 19.0

Table 3: Benchmarking results for ALICE-LP for different model and input format combinations. *ar* is the basic input configuration where only the rubric and the answers are input to the model. We use **bold** to mark the overall best results under each loss function and *underline* to mark the best-performing input format for each model. Gray underlines metrics where CL outperform SCE.

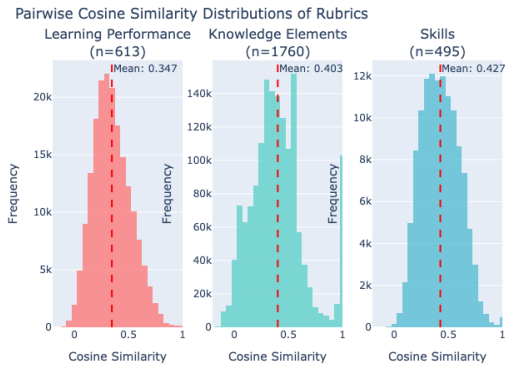

Figure 2: Distribution of pair-wise cosine similarity of rubric texts.

subtasks Table 5. Overall, rubric-based methods achieve consistently stronger performance.

The larger performance gap observed for ALICE-KE and ALICE-SK can be attributed to two factors. First, the sample solutions provided in ALICE are designed to exemplify the LP correctness. As a result, they do not reflect the full range of knowledge elements or skills. Recall in § 3 that LP is correlated with KE and SK, but they still capture distinct aspects of students' learning. Second, ALICE-KE and ALICE-SK exhibit question-dependent rubric structures with varying numbers of performance levels. This violates the fixed-label assumption of solution-based models and might exacerbate generalisation errors.

For ALICE-LP, despite a smaller performance gap, we still argue that rubric-based scoring remains competitive as it largely overcomes the level-inconsistency issue in ASAS. This is practical in real-world applications, as performance levels for each question might vary.

RQ5: Is LLM-as-encoder approach suitable for ASAS? Yes, across all the subtasks, LLMs consistently deliver improved performance over the XLM-based counterpart. The gain is particularly strong under unseen questions, suggesting a stronger generalisation ability of the LLMs. Also, larger and instruction-tuned LLMs tend to exhibit the best performance, especially on the unseen questions.

However, it is also worth noting that the gap between the LLMs and *XLM-RoBERTa*, and the gaps between LLMs of various sizes, is smaller for ALICE-KE and ALICE-SK, especially for the unseen answers. Possibly because KE and SK rubrics are similar throughout the dataset, as stated earlier.

As an external sanity check, we additionally evaluate our rubric-ranking approach on ASAP-SAS. Unlike previous methods that build ASAS models for each question, we train a single model jointly for all the questions. Since the dataset is in English and doesn't involve long context information, we benchmark it on a more diverse set of MLMs. The results reinforce the claim that LLMs-as-encoder is a promising approach for ASAS. detailed results and discussion in Appendix F.

Base Model	Input Format	ALICE-KE			ALICE-SK		
		UA (F1/QWK)	UQ (F1/QWK)	Δ(UA − UQ)	UA (F1/QWK)	UQ (F1/QWK)	Δ(UA − UQ)
XLM-RoBERTa-Long	ar	77.3/78.4	65.4/55.4	11.8/22.9	68.6/39.9	71.1/38.3	**-2.6 / 1.6**
	+q	74.1/75.9	66.2/61.7	7.9/14.1	78.0/46.7	69.7/33.7	8.3/13.1
	+s	78.1/79.6	65.9/60.7	12.2/18.9	78.2/48.2	69.2/30.9	8.9/17.3
	+qs	80.1/82.2	67.7/61.6	12.4/20.7	70.0/39.8	67.9/30.5	2.1/9.4
Llama-3.2-1B	ar	78.6/79.4	65.6/60.6	13.0/18.9	78.4/57.4	66.6/27.2	11.8/30.2
	+q	79.1/80.2	68.4/63.5	10.8/16.7	74.8/53.7	67.8/37.3	7.0/16.4
	+s	80.0/81.7	70.1/65.2	9.9/16.5	79.1/58.9	67.9/36.2	11.2/22.7
	+qs	80.7/82.5	68.4/62.9	12.4/19.7	74.8/53.3	72.6/45.3	2.3/7.9
Llama-3.2-1B-Instruct	ar	77.3/77.8	66.8/63.6	10.5/14.2	80.7/58.5	69.5/36.6	11.3/21.9
	+q	78.2/79.4	68.0/62.4	10.2/17.0	81.6/61.1	68.9/42.3	12.7/18.8
	+s	77.4/77.4	67.1/63.6	10.4/13.8	81.7/63.3	71.4/43.5	10.3/19.9
	+qs	79.8/81.0	67.9/61.2	11.9/19.8	82.5/65.3	69.3/43.3	13.1/22.0
Llama-3.2-3B	ar	79.6/82.5	70.0/64.5	9.6/18.0	80.1/56.4	71.1/42.5	9.0/13.9
	+q	77.9/79.5	69.5/65.3	8.4/14.3	80.8/60.7	70.9/46.8	9.9/13.9
	+s	77.0/76.4	70.1/64.7	6.9/**11.7**	81.2/61.6	71.4/46.1	9.8/15.5
	+qs	**80.8/83.3**	71.9/68.4	9.0/15.0	81.9/63.5	75.4/54.5	6.5/9.0
Llama-3.2-3B-Instruct	ar	76.4/76.0	67.7/64.5	8.7/11.5	81.3/58.3	71.0/46.6	10.3/11.6
	+q	78.0/78.1	66.6/61.0	11.5/17.0	**84.0/67.8**	70.5/46.4	13.4/21.4
	+s	79.0/80.0	69.7/64.9	9.4/14.9	82.5/65.7	72.3/50.9	10.3/14.9
	+qs	79.5/80.6	**73.2/67.4**	**6.3**/13.1	82.9/65.7	**75.7/57.3**	7.2/8.4

Table 4: Results of ALICE-KE and ALICE-SK with SCE. **Bold** indicates the best overall performance of each subtask. For each model, subtask, and test split, the underline indicates the best input format for each model.

Subtask	Base Model	UA (F1/QWK)	UQ (F1/QWK)	Δ(UA − UQ)
LP	*XLM-RoBERTa-Long*	67.4 / 66.2	57.3 / 49.8	10.1 / 16.4
	Llama-3.2-1B	68.9 / 69.8	58.0 / 53.6	10.9 / 16.2
	Llama-3.2-1B-Instruct	69.1 / 69.7	55.1 / 50.4	14.0 / 19.3
	Llama-3.2-3B	71.5 / 72.5	62.1 / 63.7	9.4 / 8.8
	Llama-3.2-3B-Instruct	**71.8 / 73.3**	**62.7 / 63.7**	**9.1 / 9.6**
KE	*XLM-RoBERTa-Long*	**66.6 / 51.7**	**55.8 / 37.1**	10.8 / 14.6
	Llama-3.2-1B	60.4 / 30.3	51.4 / 23.9	**9.0 / 6.4**
	Llama-3.2-1B-Instruct	59.2 / 26.3	49.9 / 9.5	9.3 / 16.8
	Llama-3.2-3B	60.1 / 15.7	52.0 / 3.3	8.1 / 12.4
	Llama-3.2-3B-Instruct	50.0 / 20.5	40.9 / 8.4	9.1 / 12.1
SK	*XLM-RoBERTa-Long*	**80.2 / 58.8**	**70.1 / 32.1**	10.2 / 26.7
	Llama-3.2-1B	75.1 / 45.6	60.6 / 25.6	14.5 / 20.0
	Llama-3.2-1B-Instruct	66.3 / 37.6	49.5 / 18.9	16.8 / 18.7
	Llama-3.2-3B	76.8 / 48.5	71.3 / 29.4	**5.5 / 19.1**
	Llama-3.2-3B-Instruct	72.4 / 42.5	60.1 / 15.7	12.3 / 26.8

Table 5: Results on ALICE benchmark with reference answer-based baseline methods across three subtasks: LP, KE and SK.

6 Conclusion

We introduce ALICE, a large-scale German ASAS benchmark that supports holistic and fine-grained assessment of student responses along three complementary dimensions: learning performance, knowledge elements, and skills. By framing ASAS as a rubric-ranking task, we provide a unified formulation that naturally accommodates ordinal, question-dependent rubric structures and extends beyond solution-based scoring.

Extensive experiments across a range of models and input configurations show that rubric-based scoring consistently improves generalisation, particularly to unseen questions. We also show that lightweight LLMs used as encoders are well-suited for ASAS. Our analysis further highlights the role of contextual information in improving robustness to question-specific distribution shifts. We hope that ALICE and our findings will facilitate more realistic and pedagogically grounded research on automated short answer scoring in NLP for Education.

7 Limitations

This work has several limitations. First, ALICE focuses on German-language STEM education, and while the dataset design is general, results may not directly transfer to other languages or subject domains. Second, although we evaluate a range of encoder architectures, we restrict our study to discriminative models and do not explore generative scoring or feedback generation. Future work on feedback generation and pedagogical capabilities of LLMs (Daheim et al., 2024; Macina et al., 2025) could benefit from integrating fine-grained assessment dimensions such as knowledge elements and skills.

While rubric-ranking proves effective and generalises well to unseen questions and question-dependent rubric structures, it requires expanding each answer into multiple answer–rubric pairs, which increases training and inference costs and limits scalability to larger models. Finally, rubric-based scoring relies on the availability of high-quality rubrics, which may not always be accessible in real-world settings. Exploring LLM-augmented rubric representations is a promising direction for future research.

8 Ethical Statement

The ALICE benchmark is constructed from student responses to assessment items collected in routine

classroom settings. We obtained all data under appropriate educational governance and privacy protections; identifiers were removed and responses were anonymised prior to annotation to safeguard student privacy. Annotators were trained educators or trained with standard calibration procedures, and we monitored for potential annotation bias across demographic groups to maintain fairness in scoring.

We acknowledge that automated short answer scoring systems can influence educational outcomes, and misuse may unfairly advantage or disadvantage students if deployed without appropriate safeguards. ALICE is released solely for research purposes under a license that prohibits high-stakes automated decision-making without human oversight. Models evaluated on this benchmark should be interpreted as tools for supporting instructional practice, not as replacements for expert human judgment. We encourage researchers to consider bias, fairness, and interpretability when developing and reporting models on ALICE, and to avoid applications that could reinforce inequitable treatment of learners.

References

Xueqing 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*, 33:992–1030.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

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.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2024. Strengths and weaknesses of automated scoring of free-text student answers. *Informatik Spektrum*, 47:78–86.

Susan M. Brookhart. 2018. *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD, Alexandria, VA.

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*, page 43–48, Berlin, Heidelberg. Springer-Verlag.

L. H. Chang, P. M. Taiga, and J. Vilén. 2024. Automatic short answer grading for finnish with chatgpt. In *Proc. of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-24)*.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. *Preprint*, arXiv:2407.09136.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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, Dublin, Ireland. Association for Computational Linguistics.

Kathleen Fisher and Joseph I. Lipson. 1986. Twenty questions about student errors. *Journal of Research in Science Teaching*, 23:783–803.

Lucas Galhardi, Bruno Brancher, Luiz Claudio Lazari, and Marco Antonio Tavares de Souza. 2018. Portuguese automatic short answer grading. In *Simpósio Brasileiro de Informática na Educação (SBIE)*. Proceedings paper; dataset for Portuguese short answer grading.

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.

Yang Han, Ziping Wan, Lu Chen, Kai Yu, and Xin Chen. 2025. From generalist to specialist: A survey of large language models for chemistry. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1106–1123, Abu Dhabi, UAE. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Saskia S. Krebs, Marijke Verbeke, Cees van der Vleuten, Filip Dochy, and Katrien Struyven. 2022. Rubrics enhance accuracy and reduce cognitive load in self-assessment and task performance. *Journal of the Learning Sciences*, 31(6):707–743.

Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Get it scored using autosas—an automated system for scoring short answers. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 9662–9669.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Zhaohui Li, Susan Lloyd, Matthew Beckman, and Rebecca 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, Singapore. Association for Computational Linguistics.

Ziyong Lin, Haoyi Wu, Shu Wang, Kewei Tu, Zilong Zheng, and Zixia Jia. 2025. Look both ways and no sink: Converting LLMs into text encoders without training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22839–22853, Vienna, Austria. Association for Computational Linguistics.

Chun Liu, Hongguang Zhang, Kainan Zhao, Xinghai Ju, and Lin Yang. 2024. LLMEmbed: Rethinking lightweight LLM's genuine function in text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7994–8004, Bangkok, Thailand. Association for Computational Linguistics.

Kun Luo, Minghao Qin, Zheng Liu, Shitao Xiao, Jun Zhao, and Kang Liu. 2024. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1354–1365, Miami, Florida, USA. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *Preprint*, arXiv:2502.18940.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

OECD-PISA. 2023. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing, Paris.

Christopher Ormerod. 2022. Short-answer scoring with ensembles of pretrained language models. *Preprint*, arXiv:2202.11558.

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

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–106, Denver, Colorado. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Gilbert Reynders, Juliette Lantz, Suzanne M. Ruder, Courtney L. Stanford, and Renée S. Cole. 2020. Rubrics to assess critical thinking and information processing in undergraduate stem courses. *International Journal of STEM Education*, 7(1):9.

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.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Are large language models good classifiers? a study on edit intent classification in scientific document revisions. *Preprint*, arXiv:2410.02028.

D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144.

Shashank Sonkar, Kangqi Ni, Lesa Tran Lu, Kristi Kincaid, John S. Hutchinson, and Richard G. Baraniuk. 2024. Automated long answer grading with ricechem dataset. *Preprint*, arXiv:2404.14316.

Judith Stanja, Wolfgang Gritz, Johannes Krugel, Anett Hoppe, and Sarah Dannemann. 2023. Formative assessment strategies for students' conceptions—the

potential of learning analytics. *British Journal of Educational Technology*, 54(1):58–75.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training BERT on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6071–6075, Hong Kong, China. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Jai Gupta, Dara Bahri, and Neil Houlsby. 2023. Transformers in the era of large language models: A survey. *arXiv preprint arXiv:2301.04655*.

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.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Fabian Zehner, Hyo Jeong Shin, Emily Kerzabi, Andrea Horbach, Sebastian Gombert, Frank Goldhammer, Torsten Zesch, and Nico Andersen. 2025. Down the cascades of omethi: Hierarchical automatic scoring in large-scale assessments. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, Vienna, Austria. Association for Computational Linguistics.

A Dataset example

See Figure 3. Note that level 1 is not present in the rubrics of knowledge elements in this example. Because level 1 is specifically for the "using without content" scenario, this level description is absent in the Physics data. Similarly, for the rubrics of skills, level 1 (partially present) is not available for this question.

Question Context Information

Prompt: Based on the patterns and measured values, derive two “the more... the more” statements for the alignment of the PV cell.

Sample Solution:
The more perpendicular the PV cell is aligned to the radiation source, the higher the electrical energy. The more light reaches the PV cell (the less the PV cell is shaded), the higher the electrical energy.

Student Response

Answer: The more directly the light source shines on the solar panel, the higher the generated voltage. The more of the panel was covered, the less voltage was generated.

Learning Performance

Level 0: The students do not formulate any "the more - the more" statements regarding the angle of incidence and the illuminated area..

Level 1: Students formulate one statement about angle or illuminated area.

Level 2 : Students formulate two statements about angle *and* illuminated area.

Knowledge Elements

Electrical Energy
Level 0: The term or similar meanings are not mentioned..
Level 2: Non-targeted use: The students use the voltage to explain something other than energy conversion.
Level 3: Targeted use: voltage is used to make statements about energy conversion.

Radiation Energy
Level 0:The term or similar meanings are not mentioned..
Level 2: Non-targeted use: The students use angles or shading to explain something other than energy conversion.
Level 3: Targeted use: angle or shading is used to make statements about energy conversion.

Conversion
Level 0: The term or similar meanings are not mentioned..
Level 2: Non-targeted use: The students use transformation concepts to explain something other than the conversion of radiant energy into electrical energy.
Level 3: Targeted use: student indicates that radiation is converted to electrical energy.

Skill

Reasoning
Level 0: Not present: No explanation using experimental results.
Level 2 : Present: Explains patterns using experimental results; includes at least one “the more... the more” sentence.

Figure 3: Illustrative instance from the short-answer scoring benchmark, translated from German to English. Selected rubric levels shown in green.

B Input formats of LLMs

In Ruan et al. (2024), they experimented with two input formats for LLM-based classification: *natural language* and *structured*. The former uses natural language as sequence boundaries, and the latter marks sequence boundaries with XML-style markup. Below is an example input text for the LLMs. The *structured* formatting approach achieved better results. We adopt this strategy in our paper:

```
<Frage>Beschreibe auf Basis Deines
↪ Vorwissens die Erklärungskraft der
↪ submikroskopischen Ebene (was vermag
↪ diese Modellebene
↪ darzustellen?).</Frage>
<Antwort>Die submikroskopische Ebene
↪ stellt Strukturformeln von Molekülen
↪ modellhaft dar.</Antwort>
<Rubrik>Die Schüler:innen erläutern den
↪ erweiterten Blick der
↪ makroskopischen Ebene
↪ nicht.</Rubrik>
```

They also experimented with prepending natural language instruction to the input and found that it has a limited impact on the final performance.

C Dataset collection, annotation and further statistics

Dataset collection, annotation The students' answers are collected from Moodle courses used in high schools in **Blinded** state, Germany. The Moodle courses cover four subjects: biology, chemistry, mathematics and physics. The ALICE-SK and ALICE-SK don't include rubrics for mathematics. The dataset was scored in four phases, each for a subset of the contained questions. Each phase was further grouped into a pilot phase and an annotation phase. The annotation was performed on the INCEPTION (Klie et al., 2018) platform. In the pilot phase, for each domain and question, the annotators were trained to score answers using a smaller subset of the data until a desirable Cohen's $\kappa > 0.75$ was reached per question. Where needed, initial scoring rubrics and question-wise guidelines were revised for better clarity, and the annotators were retrained using the updated guidelines. Following this, the remaining student answers were distributed among the different annotators. Due to the size of the overall dataset, multiple annotators needed to be replaced during the annotation process, resulting in minor fluctuations across the four phases.

For learning performance, we reached a Cohen's Kappa score of 72; For knowledge element, the Cohen's Kappa score is 81. And for skills, the score is 80.

Further statistics See Table 6 for per-subject statistics and Figure 4 for level distribution across subtasks.

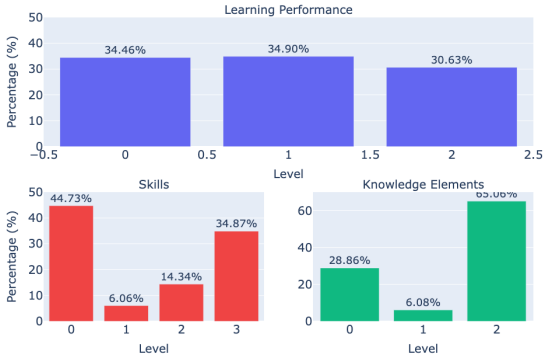

Figure 4: Distribution of levels across the subtasks

Subject	answers	questions	ke	sk
Biology	1987	11	8	11
Chemistry	8977	66	70	23
Mathematics	1257	14	NA	NA
Physics	4353	21	8	12

Table 6: Per-subject number of answers, questions, knowledge elements and skills.

D Model configuration and Hyperparameters

Model Selection Rationale. XLM-Roberta-Long [3] is chosen as a strong multilingual encoder baseline with extended context length, which is important for handling multi-sequence inputs consisting of answers, rubrics, questions, and sample solutions.

In addition, we evaluate a set of lightweight LLMs with parameter sizes between 1B and 3B. These models offer substantially larger pretraining corpora and stronger reasoning capabilities than MLMs, while remaining feasible for fine-tuning under realistic computational constraints. We include both base and instruction-tuned variants to

[3] markussagen/xlm-roberta-longformer-base-4096

examine whether instruction tuning confers advantages in discriminative rubric-ranking settings.

For MLM-based encoders, we use the [CLS] token representation as the sequence embedding. For LLM-based encoders, we extract the final [EOS] token representation. In both cases, the sequence embedding is passed through a feedforward layer to produce alignment scores.

Reproducibility Considerations. All experiments are conducted using fixed random seeds. Model checkpoints, preprocessing scripts, and evaluation code will be released upon publication to facilitate reproducibility and future benchmarking.

We randomly sample 10% of the training data as the validation set and evaluate the model on it at each epoch. The model checkpoint that achieves the highest accuracy is saved for testing.

MLM-based models We fine-tune with a learning rate of 2×10^{-5} for 8 epochs with a batch size of 16.

LLM-based models We train with 4-bit quantisation (QLoRA), a LoRA rank of 64, and a learning rate of 1×10^{-4}. We use a batch size of 4 and gradient accumulation of 8 steps. For ALICE-LP, we train the model for 6 epochs. During the experiment, we found that the model converges after 4 epochs on ALICE-KE and ALICE-SK, so we train these subtasks with 4 epochs on LLMs.

E Benchmarking with Close-source LLM

E.1 Prompt

E.1.1 System Prompt

Listing 1: Base System Prompt (German)

```
Sie werden die Antwort eines K12-Schuelers
    in MINT-Faechern bewerten.

ZIEL
Weisen Sie basierend auf dem
    bereitgestellten Rubrikensatz genau EINE
     ganze Zahl als Punktzahl zu. Sie
    muessen die Schuelerantwort intern
    anhand jeder Rubrik bewerten und dann
    die am besten passende Punktzahl
    auswaehlen.

FOKUS: [Task-specific focus, see below]

REGELN
- Beruecksichtigen Sie die Frage und die
    Beispielantworten nur, wenn sie
    bereitgestellt werden; sie sind
    optionale Hilfsmittel zum Verstaendnis,
    keine strikten Anforderungen.
- Eine Rubrik ist nur erfuellt, wenn ihre
    erforderlichen Kriterien erfuellt sind.
    Wenn eine Rubrik "eines von/von mehreren
    " verwendet, folgen Sie dieser Logik.
- Bewerten Sie ausschliesslich basierend auf
     der Uebereinstimmung zwischen der
    Schuelerantwort und den Rubrikkriterien
    (und der Frage, falls vorhanden).
    Ignorieren Sie den Stil, es sei denn, er
     wird von der Rubrik verlangt.
- Geben Sie nur eine ganze Zahl zurueck, die
     der hoechsten erfuellten Rubrik
    entspricht.

AUSGABE
Geben Sie NUR ein gueltiges JSON-Objekt mit
    genau diesen Feldern zurueck:
{
  "reasoning": "1-3 praegnante Saetze, die
    erklaeren, warum diese Punktzahl am
    besten zur Rubrik passt.",
  "score": <Ganzzahl>
}

WICHTIG
- Fuegen Sie KEINE zusaetzlichen Schluessel,
     Texte oder Formatierungen ausserhalb
    des JSON hinzu.
```

E.1.2 Task-Specific Prompt

Listing 2: Learning Performance (LP)

```
FOKUS: Bewerten Sie die GESAMTLEISTUNG des
    Schuelers bei der Bearbeitung der
    Aufgabe, einschliesslich Verstaendnis,
    Anwendung und Kommunikation des Wissens.
```

Listing 3: Knowledge Elements (KE)

```
FOKUS: Bewerten Sie die BEHERRSCHUNG
    SPEZIFISCHER WISSENSELEMENTE (z.B.
    chemische Reaktionen, physikalische
    Gesetze, mathematische Konzepte).
    Konzentrieren Sie sich auf die
    Korrektheit und Tiefe des fachlichen
    Verstaendnisses der jeweiligen
    Wissenskomponente.
```

Listing 4: Skills (SK)

```
FOKUS: Bewerten Sie die ANWENDUNG
    SPEZIFISCHER FAEHIGKEITEN (z.B.
    Hypothesen formulieren, Experimente
    entwerfen, Daten interpretieren,
    Schlussfolgerungen ziehen).
    Konzentrieren Sie sich auf die Qualitaet
     der methodischen und prozeduralen
    Fertigkeiten, nicht nur auf das
    Fachwissen.
```

E.1.3 User Prompt

Listing 5: User Prompt

```
RUBRIKENSATZ
{rubrics}

FRAGE
```

```
   {question}

   BEISPIELLOESUNGEN
   {sample_solutions}

   SCHUELERANTWORT
   {answer}
```

E.2 Results

Base Model	Input Format	LP		KE		SK	
		UA (F1/QWK)	UQ (F1/QWK)	UA (F1/QWK)	UQ (F1/QWK)	UA (F1/QWK)	UQ (F1/QWK)
GPT-4o-mini	*ar*	59.1 / 54.9	60.2 / 53.6	52.2 / 53.5	51.2 / 54.0	65.9 / 36.1	63.1 / 41.2
	+q	60.5 / 57.3	60.7 / 57.6	47.7 / 51.9	47.6 / 52.4	70.7 / 43.6	66.7 / 43.8
	+s	58.0 / 53.4	60.0 / 53.6	52.8 / 53.9	52.4 / 54.1	66.3 / 36.6	62.7 / 41.0
	+qs	60.8 / 56.9	60.5 / 58.9	48.4 / 51.7	47.8 / 52.6	70.2 / 42.5	66.6 / 44.3

Table 7: Prompt-based results on ALICE benchmark. F1 and QWK scores are reported as percentages.

F Benchmarking on ASAS-SAP

Table 8: Per-question QWK scores on ASAP-SAS dataset

Joint Rubric-Ranking											
Base Model	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Mean
BERT-base-cased	81.5	77.4	67.6	70.9	78.6	81.9	67.6	62.9	78.1	71.4	73.8
RoBERTa-base	83.3	72.0	68.9	76.0	77.2	83.2	70.3	69.4	79.8	75.5	75.5
ModernBERT-base	76.3	57.5	68.3	65.7	73.4	75.6	56.8	52.5	79.6	70.2	67.6
XLM-RoBERTa-Long	80.9	70.8	69.8	73.1	82.7	84.2	67.7	56.8	79.2	74.4	73.9
Llama-3.2-1B	84.6	80.9	69.5	77.0	81.0	80.7	72.0	67.7	82.8	76.3	77.2
Llama-3.2-1B-Instruct	81.5	77.8	68.5	74.7	82.9	83.5	69.1	68.9	82.0	76.4	76.5
Llama-3.2-3B	85.8	80.7	71.3	69.5	83.7	86.3	71.6	71.3	83.9	77.1	78.1
Llama-3.2-3B-Instruct	87.8	83.5	67.4	69.2	82.9	85.4	71.4	72.1	83.1	79.2	78.2
Previous Methods											
Method	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Mean
Ramachandran et al. (2015)	86.0	78.0	66.0	70.0	84.0	88.0	66.0	63.0	84.0	79.0	78.0
Riordan et al. (2017)	79.5	71.8	68.4	70.0	83.0	79.0	64.8	55.4	77.7	73.5	72.3
Wang et al. (2019)	79.2	71.4	NA	NA	80.4	79.3	NA	NA	NA	75.3	NA
Kumar et al. (2019)	87.2	82.4	745	743	845	858	725	624	843	832	791
Bexte et al. (2022) MAX	84.0	79.0	67.0	67.0	72.0	58.0	68.0	77.0	72.0	72.0	72.4
Bexte et al. (2022) AVG	88.0	78.0	74.0	72.0	69.0	74.0	58.0	69.0	77.0	72.0	73.1

We also benchmarked our method on ASAP-SAS, which contains diverse domains such as English Literature, and Science Bexte et al. (2022). Note that all the listed methods train a separate model for each question. Below are short descriptions of each method.

Ramachandran et al. (2015) The paper proposes a supervised short-answer scoring approach that represents student responses using a combination of lexical overlap, semantic similarity (via WordNet and distributional semantics), and question-specific reference answers, which are then fed into a regression/classification model. The method is deliberately feature-driven rather than neural, reflecting the low-resource and interpretability constraints of the time.

Kumar et al. (2019) The paper proposes AutoSAS, a supervised short-answer scoring system that predicts a numeric score for a student response using engineered linguistic and semantic features. For each answer, the system extracts lexical diversity measures, Word2Vec embeddings, and overlap features between the student response and the prompt or reference content. These features are then fed into a regression model that outputs a score. They built a separate model for each question.

Wang et al. (2019) It proposes a neural short answer grading model that augments a baseline encoder with an explicit rubric component. The architecture has two parts: (1) a base component that encodes student answers using a BiLSTM over word embeddings to produce a feature vector, and (2) a rubric component that computes word-level attention alignments between the answer and each key element in the rubric to produce rubric-aware features. These features are merged (by concatenation or weighted sum) and passed through a regression layer to predict the score, allowing the model to learn alignment between rubric criteria and answer content.

Ormerod (2022) The authors fine-tune a variety of transformer-based pretrained language models (small, base, large) on the Kaggle Automated Short Answer Scoring dataset (ASAS) task, train a separate feature-based model, and then build ensembles of these models (via logistic regression over model log-probabilities) to produce the final score. They built a separate model for each question.

Riordan et al. (2017) The paper evaluates simple neural architectures for short-answer scoring by training CNN- and LSTM-based models on student responses, using word-embedding inputs and treating scoring as regression or classification depending on the dataset. The models take only the

student's answer text (plus prompt-specific training data) and do not incorporate rubrics or reference criteria as structured inputs. They compare these neural models to a strong feature-engineered baseline across three SAS datasets and find that neural models can outperform the baseline, though the best architecture varies with prompt characteristics. They built a separate model for each question.

Bexte et al. (2022) The authors train the model, which, given an answer pair, predicts if it belongs to the same level. During testing, the model compares the candidate answer with answers of various levels to determine the final level. During testing, they used the MAX or AVG similarity between the candidate's answer and anchor samples to determine the final score.

Table 8 reports per-question QWK scores on ASAP-SAS using the standard per-prompt evaluation protocol. Overall, our joint rubric-ranking models achieve competitive but **not state-of-the-art** performance compared to prior work that trains separate models for each prompt. In particular, while some previous methods attain higher peak scores on individual questions, they rely on prompt-specific training and tuning, which deviates from realistic deployment scenarios where new questions are encountered without retraining.

In contrast, our approach uses a single jointly trained model across all prompts. Among our models, lightweight LLM encoders (e.g., Llama-3.2-3B variants) outperform MLM baselines and achieve mean QWK comparable to or exceeding most prior neural approaches, despite the more challenging joint-training setting. These results indicate that rubric-ranking with joint training trades off maximal per-prompt performance for improved generality and practical applicability, aligning better with real-world ASAS use cases.