Are Rubrics All You Need? Towards Flexible Rubric-based Automatic Short-Answer Scoring via Attention-based Span Alignment and Pairwise Ranking

Anonymous ACL submission

Abstract

In educational assessment, scoring rubrics are essential to the practitioner's toolbox since they define the exact criteria for scoring learner responses. However, in past NLP research on automatic short-answer scoring, scoring rubrics are rarely used as explicit scoring references and, if used, mostly treated as supplementary input. With this study, we aim to explore different possible implementations for rubric-based short answer scoring where models are explicitly conditioned towards using a provided rubric as a scoring reference. For this purpose, we propose GRASP, a novel pointer-based architecture that uses bilinear attention to predict the alignment between pooled span embeddings of student answers and rubric criteria from a single encoder forward pass. Moreover, we explore SBERT and Cross Encoders for pairwise ranking, and include five-shot prompting generative LLMs as baseline. We compare all methods using a novel German short answer scoring dataset and the established English ASAP-SAS. Results reveal that the effectiveness of the different methods depends on the nature of the dataset. For ASAP-SAS, pairwise ranking achieves a competitive performance close to the state of the art, while GRASP underperforms. However, for the German dataset, this is reversed. There, GRASP significantly outperforms the other methods and generalises better to unseen questions.

1 Introduction

Automatic short-answer scoring (ASAS), also known as automatic short-answer grading (ASAG), short-answer assessment (ASAA) or constructed response assessment, refers to a set of techniques for automatically scoring student answers within tests or other assessment forms, such as those provided by intelligent tutoring systems (Bai and Stede, 2023; Bexte et al., 2024; Burrows et al., 2015). Modern short-answer scoring systems are mainly built in two abstract setups. In the first established setup, learner responses are directly classified by ML models with corresponding scores as labels, respectively, as regression targets, which Bexte et al. (2022) refer to as *instance-based*. In the second setup, student answers are compared to reference answers or representative clusters of those (Bexte et al., 2022; Zehner et al., 2016). This is implemented with the help of either transformer-based language models trained in a natural language inference setup (Sung et al., 2019; Camus and Filighera, 2020) or various methods relying on embedding spaces and vector arithmetic (Bexte et al., 2022; Zehner et al., 2016), which Bexte et al. (2022) fittingly refers to as *similarity-based*.

However, both approaches slightly differ from human assessors' typical method of scoring student responses. In the majority of cases, human assessors rely on scoring rubrics, documents that outline the criteria a response must meet to receive a specific score (Reddy and Andrade, 2010). When using instance-based scoring, models must infer scoring criteria during training and establish a latent form of these within their internal representations. This usually limits the models to questions seen during training. By contrast, similarity-based scoring allows for a more generalised use of the resulting models and even a certain degree of domain transfer capabilities (Sung et al., 2019; Camus and Filighera, 2020). However, such models are still limited to cases where the quality of a response can be measured by its similarity to a given reference answer, and not all types of assessment questions necessarily fit into this category, as, for example, Bloom's taxonomy suggests (Bloom et al., 1956).

The reliance on sample solutions for giving scoring models domain transfer capabilities stems from days when natural language processing relied mostly on linear models and feature engineering. In traditional feature engineering, differences between a reference answer and a student response are more straightforward to represent than whether

a student response complies with the criteria defined in a rubric (Burrows et al., 2015). However, transformer-based language models (Vaswani et al., 2017) have demonstrated remarkable language understanding capabilities, which leads to the question of whether these models can successfully understand rubrics and learn how to use them as primary scoring references.

- With this paper, we provide a formal definition for rubric-based short answer scoring.
- We introduce GRASP (Guided Rubric Alignment with Span Pooling). This encoder-based architecture frames rubric-based short answer scoring as a span alignment task. Unlike prior work using rubrics as auxiliary features (e.g., Li et al. 2023a), GRASP treats rubrics as dynamic span-aligned scoring anchors, enabling true rubric-conditioned scoring and transfer.
- We also implement the task as a pairwise ranking problem using task-specific fine-tunes of the well-known *SBERT* and *Cross Encoder* architectures (Reimers and Gurevych, 2019).
- We introduce the *ALICE-LP* dataset, a novel large-scale German-language dataset specifically aimed at rubric-based short answer scoring, which was collected in German middle- and high schools in the context of STEM education units.
- We also provide secondary evaluations of our systems on the public *ASAP-SAS* dataset.
- Our findings suggest that the performance of the different approaches seems to be dataset-dependent, with *GRASP* excelling for *ALICE-LP*, demonstrating the best transfer capabilities to unseen questions, while falling short for *ASAP-SAS*. On the other hand, we observed a reversed pattern for *SBERT* and *Cross Encoders*.

2 Background

2.1 Rubrics

Rubrics define scoring rules for one or multiple scoring dimensions used to score responses in the context of various assessment forms (Reddy and Andrade, 2010; Panadero and Jonsson, 2013). Usually, a rubric defines for a given question what criteria a text response must meet to be considered evidence for a certain performance level (e.g., full, partial, or no credit). They are widely used in primary, secondary, and post-secondary education and have become the de facto standard in performance assessment and standardised testing (Panadero and Jonsson, 2013). The primary purpose of rubrics is to provide coherent criteria that help practitioners systematically assess learners' responses (Reddy and Andrade, 2010). In the context of Evidence-Centered Design (Mislevy et al., 2003), a popular conceptual framework for implementing assessments that consider the validity of score interpretations from the beginning, rubrics play a central role since they define exactly what evidence looks like for the respective performance level. Table 1 shows an example rubric from the *ALICE-LP* dataset.

2.2 Automatic Short Answer Scoring

Automatic short-answer scoring refers to the task of automatically assigning a short written response of up to one paragraph a corresponding performance level, signalling the quality of that response according to task-specific scoring criteria (Bai and Stede, 2023; Bexte et al., 2024; Burrows et al., 2015). Over the past decades, this has been approached in various ways (Burrows et al., 2015; Bai and Stede, 2023). Older approaches often measure the amount of word overlap (Siddiqi and Harrison, 2008; Cutrone and Chang, 2010) or look at the overlap on the level of various syntactic or semantic formalisms between a sample solution and a given response (Bachman et al., 2002; Mitchell et al., 2002; Hahn and Meurers, 2012). Another approach relies on assigning scores based on similarity of a response to representative clusters from a given training set (Zehner et al., 2016). Newer approaches rely on multiple forms of embedding spaces to compare responses with sample solutions (Bexte et al., 2022), or various text classifiers aimed at classifying student responses or pairs of responses and reference answers (Ramachandran et al., 2015; Saha et al., 2018; Riordan et al., 2017; Sung et al., 2019; Kumar et al., 2019; Camus and Filighera, 2020; Ormerod, 2022; Gombert et al., 2023; Li et al., 2021; Filighera et al., 2022; Padó et al., 2023). An established approach is to fine-tune transformer encoder language models to classify either tuples of responses and sample solutions or triples of the same with the question text added as a third element (Sung et al., 2019; Camus and Filighera, 2020; Fernandez et al., 2022; Gombert et al., 2023).

Level	Criteria
0	The students do not describe the conversion (into thermal energy/heat).
1	The students use the term conversion without addressing thermal energy, or mention thermal energy in connection with energy loss.
2	The students describe that the remaining energy was converted to thermal energy/heat.

Table 1: an example rubric depicting possible performance levels and the corresponding criteria for a task on energy conversion.

Also, there has been ongoing research on systems that promote semiautomatic scoring where human assessors are supported by automated systems comparing responses to sample solutions (Li et al., 2023b; Andersen et al., 2023). Moreover, Kortemeyer (2024) explored the degree to which zero-shot prompting GPT-4 with a hand-crafted prompt can solve the problem of automatic short answer scoring, with the result that the model performed significantly worse than most transformer-encoder-based approaches relying on fine-tuning such as the ones proposed by Camus and Filighera (2020) or Sung et al. (2019).

These results were also further confirmed by a more systematic study by Ferreira Mello et al. (2025), who compared a range of feature-based and neural models to various LLM prompting techniques and came to the conclusion that even older feature-based models outperformed zero-shot prompting in nearly all cases. Results by Chamieh et al. (2024) paint a similar picture. Compared to this, Claude Sonnet, when paired with retrieval-augmented generation, seems to achieve a more desirable zero-shot performance similar to or better than previously published results achieved by smaller fine-tuned models (Wang and Ormerod, 2024).

2.3 Automatic Short Answer Scoring using Rubrics

Scoring rubrics have already been explored as model input in various past works. Marvaniya et al. (2018) implemented a system that uses scoring rubrics as input to feature-based scoring systems. However, what they call scoring rubrics are effectively collections of reference answers representative of clusters of similar answers encountered in the training set for a given task, instead of a description of individual scoring criteria, and, thus, their approach is rather similarity-based. Wang et al. (2019) implemented a BiLSTM-based system that computes word-level alignment scores between words in a student's answer and those in a corresponding rubric. They find that their model outperforms a regular BiLSTM-based scoring model published by Riordan et al. (2017) in low-resource settings, thus providing substantial evidence for the potential of scoring rubrics in short answer scoring.

Li et al. (2023a) proposed an architecture based on a recurrent relational network fed with BERT embeddings of responses, reference answers, question texts, and scoring rubrics if the latter are contained in a given dataset. However, their work does not specifically focus on how scoring rubrics influence the overall outcome. Sonkar et al. (2024) evaluated multiple models using rubrics for a closely related task setup called "long answer scoring," i.e., the scoring of answers that consist of a few paragraphs of text without being complete essays. For this purpose, they successfully fine-tuned various established transformer encoders in a natural language inference setup and prompted various foundational LLMs. Wei et al. (2025) found that rubric-based prompting of LLMs surpasses the performance of regular few-shot learning and can, moreover, benefit data synthesis of artificial short answer scoring data.

3 Method

In general, scoring rubrics can vary strongly from assessment to assessment and question to question. They vary in the number of performance levels they distinguish and the exact criteria for assigning them. They might be formulated holistically or include multiple scoring dimensions that are scored individually. Moreover, for some questions, rubric definitions corresponding to higher scores might semantically entail definitions for lower scores and add upon those, but this is not the case for all rubrics, since there can be cases where particular performance levels correspond to distinct criteria without any overlap. To flexibly address such diverse cases, the architecture must satisfy the following properties:

1. It must accommodate rubrics with varying dimensionality and number of performance levels.

2. It must allow for generalisation towards unseen rubrics if trained on sufficiently enough data, so the resulting models can, in theory, be applied to unseen questions and domains.

An implementation that supports these features is to interpret the task as a ranking problem. Put formally, let:

- A be a student answer (open-ended text),
- $\mathcal{R}_q = \{r_1, r_2, \ldots, r_n\}$ be the set of all performance level-wise text spans associated with a given scoring rubric for the question q. Each $r_i \in \mathcal{R}_q$ is a textual description of the rubric criteria that must be met to assign a performance level n for a given scoring dimension.

The goal is to learn a retrieval function

$$f : A \times \mathcal{R}_q \to \mathbb{R} \tag{1}$$

that assigns a alignment scores between all $r_0, ..., r_n$ and the given student answer.

At inference time, f induces a ranking,

$$R' = \text{sort}(r \in \mathcal{R}_q, \text{ by } f(A, r)), \tag{2}$$

and the performance level associated with the highest ranked r_n is selected.

3.1 GRASP: Guided Rubric Alignment with Span Pooling

First, we introduce *GRASP*, ***G**uided **R**ubric **A**lignment with **S**pan **P**ooling*, an encoder-based architecture. Inspired by the Pointer Networks architecture (Vinyals et al., 2015) and pre-transformer age models for span/section alignment using bilinear attention (Chen et al., 2016; Lu et al., 2016), we model the rubric selection task as a span alignment problem over a variable-length input set. GRASP treats rubric spans as dynamic scoring anchors and computes direct attention-based alignment in a unified encoder pass. To our knowledge, this is the first model to operationalise short-answer scoring as rubric-level span selection.

GRASP can be implemented with any transformer encoder language model. This model receives a given student answer, the corresponding question, its context (if given), and all rubric descriptions in a single forward pass. Using bilinear attention (Lu et al., 2016), the model computes the attention between the student answer span and each rubric span, allowing it to assess contextual relevance and effectively map the answer to the most relevant rubric. The rubric with the highest alignment score is then selected, and the corresponding performance level is assigned to the student's answer, akin to pointer networks choosing output positions (Vinyals et al., 2015).

Compared to dual-encoder models such as ColBERT (Khattab and Zaharia, 2020), our approach employs full cross-attention between the answer and all rubric spans, enabling deeper interaction and richer contextual relevance modelling. This design facilitates flexible handling of rubrics with varying performance levels and supports input-conditioned prediction, allowing the model to adapt to diverse evaluation criteria.

Starting with a transformer encoder model, let $\mathbf{H} \in \mathbb{R}^{T \times d}$ be the matrix of contextualized token embeddings obtained from this model over the input sequence x of length T, with hidden size d:

$$\mathbf{H} = \text{Encoder}(x) \tag{3}$$

$\mathbf{H}$ contains student answer, question text (plus additional context, if given), and the performance level-wise criteria. Each performance-level wise description of criteria $r_i \in \mathcal{R}_q$ corresponds to a token span $[b_i, e_i)$, and its representation $\mathbf{r}_i$ is computed via mean pooling:

$$\mathbf{r}_i = \frac{1}{e_i - b_i} \sum_{t=b_i}^{e_i-1} \mathbf{H}_t \quad \text{for } i = 1, \ldots, N \tag{4}$$

The student answer also corresponds to a span $[b_a, e_a)$ and is pooled similarly:

$$\mathbf{a} = \frac{1}{e_a - b_a} \sum_{t=b_a}^{e_a-1} \mathbf{H}_t \tag{5}$$

Each rubric receives a relevance score α_i relative to the answer using a bilinear attention function:

$$\alpha_i = \mathbf{r}_i^\top \mathbf{W} \mathbf{a} \tag{6}$$

We then calculate a softmax over all α_i:

$$p_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{N} \exp(\alpha_j)} \tag{7}$$

Finally, the model is trained using cross-entropy loss with the target rubric index $y \in \{0, ..., N-1\}$:

$$\mathcal{L} = -\log p_y \tag{8}$$

During inference, we then determine the performance level (i.e., score) assigned to the student answer using argmax:

$$\hat{i} = \arg\max_{i=1,\ldots,N} p_i \quad \Rightarrow \quad \hat{s} = \text{Score}(r_{\hat{i}}) \tag{9}$$

Figure 1 depicts this architecture.

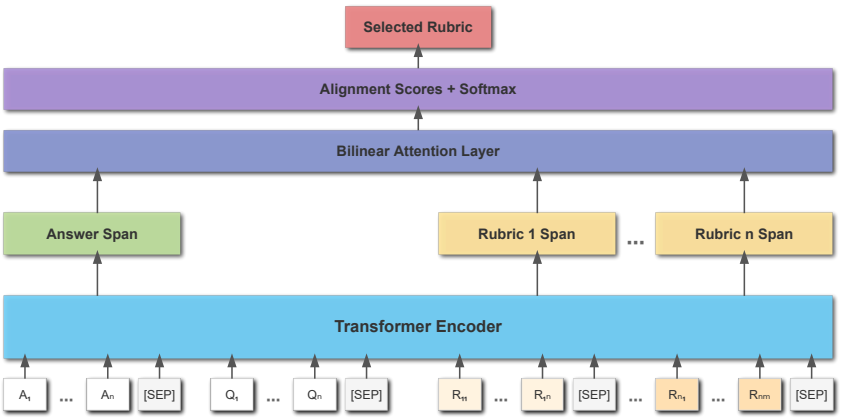

Figure 1: This figure depicts GRASP, our novel pointer-style architecture for rubric-conditioned scoring. The model aligns student answer spans to rubric criteria spans in a single encoder pass via bilinear attention. $A_1, ...A_n$ refers to the tokens of the student answer, $Q_1, ..., Q_n$ to the tokens of the question text, and $R_{1,1}, ..., R_{m,n}$, refers to the tokens of the different rubric spans. For multi-label settings, e.g., when using *GRASP* with analytic rubrics, *softmax* might be replaced with *sigmoid*.

3.2 Pairwise Ranking via SBERT

An established approach for solving ranking problems (e.g., in information retrieval) is the use of *sentence embeddings* (Reimers and Gurevych, 2019). In the context of*ASAG* Ranking is conducted by embedding a student's answer and the rubric criteria of each performance level individually into a shared vector space to acquire embeddings for both, whose distance we can then measure. During inference, a final performance level is assigned by ranking the different rubric embeddings by their similarity to the student answer embedding and selecting the best-ranked. To tackle our problem in this way, we use regular *cosine similarity loss* as the distance metric.

To train the model, we restructure the given training set so it consists of pairs of student answers and rubric criteria with a corresponding target similarity score that is set to 1 in case of compliance and to 0, otherwise. Question spans are included in both such inputs to provide context. These are added after the answer, respectively, rubric criteria, separated by a separation token. During inference, we determine similarity using pairwise comparisons and rank by the achieved cosine similarity. Compared to *GRASP*, this requires multiple forward passes.

3.3 Pairwise Ranking via Cross-Encoder

The second established approach stemming from the context of *SBERT* (Reimers and Gurevych, 2019) is the usage of cross-encoders. Cross-encoders are fine-tuned transformer encoders that directly predict a distance score for two given input sentences, unlike *SBERT*, where a comparison is carried out between embeddings generated by an encoder using vector arithmetic. By rule of thumb, *cross-encoders* are assumed to be more precise in their comparison (Bexte et al., 2022). To fine-tune cross-encoders, we convert the dataset in the same way as we did for *SBERT*. Questions are included here as well (however, here, only once, since similarity is of course computed within a single forward pass) to provide context. We then train the model using *Binary Cross-Entropy Loss* to classify whether a student answer complies with a the criteria for a performance level. We then rank by the resulting confidence scores.

3.4 Prompting LLMs

In line with work as conducted by Kortemeyer (2024), Ferreira Mello et al. (2025), and Wang and Ormerod (2024), we also evaluated prompting LLMs (for the exact prompts, see Appendix) for our purpose. In this context, we focus on five-shot prediction. Accordingly, examples for the five-shot were randomly sampled from the training data.

4 Experiments

4.1 Datasets

4.1.1 ALICE-LP

This is a novel German language dataset we devised for training and evaluating rubric-based automated short-answer scoring systems. Its name stands for BLINDED. We collected this dataset at German middle and high schools (*Gemeinschaftsschule* and

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	1/2

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

Set	#Questions	#Answers	#Levels (0/1/2)
Train	40	10,317	4,981/3,222/2,114
UA	40	1,167	563/363/241
UQ	15	3,924	2,245/1,060/619

Table 3: Distributions in the ALICE-LP dataset.

Set	#Questions	#Answers	#Levels (0/1/2/3)
Train	10	17,043	6,731/5,579/3,992/741
Test	10	5,024	2,053/1,590/1,329/252

Table 4: Distributional properties of the ASAP-SAS dataset.

Gymnasium) in the state of BLINDED. All parents of the students whose data we used signed a data sharing and reuse agreement and data from students whose parents did not sign the agreement was excluded from the dataset. The data collection was permitted by the ethics committee of the BLINDED institute.

The dataset includes answers from three domains: biology, mathematics, and physics to an overall of 55 questions. All questions within the dataset follow the paradigm of evidence-centred design (Mislevy et al., 2003) and aim at formatively assessing students' overall learning progression (Kubsch et al., 2022) within Moodle courses while simultaneously acting as learning activities. Each question was devised by didactics experts who also implemented the Moodle courses and comes with detailed scoring guidelines, including scoring rubrics.

The dataset was scored using the INCEpTION annotation tool (Klie et al., 2018). All annotators were student assistants enrolled in a teacher education programme and familiar with the covered topics. The dataset was scored in four phases, each dealing with a subset of the contained questions. Each phase was further grouped into a pilot phase and an annotation phase. In the pilot phase, for each domain and individual question, the corresponding human 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 the present study, the dataset is divided into a training set, which contains 40 different questions, and two test sets, namely *unseen answers*, containing 20% of answers to each questions of the 40 seen during training sampled with stratification, and *unseen questions*, containing 15 questions from each domain which were not present in the training set. This setup was inspired by the *SciEntsBank* (Dzikovska et al., 2013) and *Short Answer Feedback* (Filighera et al., 2022) datasets (which we did not consider for this work since they do not include rubrics), and allows us to separately assess how well models perform for seen in-domain questions, and how well they can transfer their knowledge to unseen questions. Table 3 shows the overall distributional properties of the dataset.

4.1.2 ASAP-SAS

ASAP-SAS is a widely used benchmark dataset for short answer scoring released initially in the context of a Kaggle competition[1]. Besides being one of the most established benchmarks in automatic short answer scoring, we primarily included this dataset since it comes with rubrics, which are provided

[1]https://www.kaggle.com/competitions/asap-sas/data

Model	UA	UQ
GBERT-large GRASP	86.68†	**81.32**†
GELECTRA-large GRASP	**87.80**†	80.09†
GELECTRA-large CrossEnc	83.79	77.34
GBERT-large SBERT	85.82	77.14
GELECTRA-large SBERT	85.77	77.08
GBERT-large CrossEnc	81.44	75.12
GPT-4o 5-shot	61.80	63.80
Qwen 3-8B 5-shot	55.51	60.57
GPT-3.5 Turbo 5-shot	50.44	54.76
Mistral-7b-Instruct-v0.3 5-shot	47.59	53.63
Llama-3.1-8B-Instruct 5-shot	49.15	51.66

Table 5: Best weighted F1 scores achieved by the different assessed approaches for the three ALICE-LP test sets. UA = unseen answers. UQ = unseen questions. †*Significantly better than the SBERT and Cross Encoder models as well as the LLMs ($p < .001$, randomisation test).*

Model	$\text{Mean}_{\text{Fisher}}$ QWK
ModernBERT-large GRASP	70.09
ModernBERT-large SBERT	76.96
ModernBERT-large CrossEnc	77.95
Ramachandran et al. (2015)	*77.87**
Riordan et al. (2019)	*77.88*
Kumar et al. (2019)	*80.15**
Ormerod (2022)	***80.61****
Human Agreement	*90.30*

Table 6: Best Fisher-weighted mean Quadratic Weighted Kapppa scores achieved by the different assessed approaches for the ASAP-SAS test set. Additionally, we list the best published results for this dataset. In the appendix, we included brief descriptions of these baselines. **Calculated based on the task-wise results the authors report. Their papers only report regular means and individual task results.*

as supplementary material. It consists of answers to 10 questions covering various topics, including STEM and reading comprehension, and was collected from US-based high schools. Questions are scored on a range from 0 to 2 or 0 to 3 and reflected in both the train and test sets. All reading comprehension questions also supply the corresponding texts as context. Where present, we included these as part of the question spans. The established way of evaluating systems using this dataset is to calculate item-wise quadratic weighted kappa scores and use those to calculate a Fisher-normalised mean for the overall dataset. Scores from a second human assessor are provided so that human agreement can be calculated and compared to systems evaluated with this dataset. Table 4 shows the distributional properties.

4.2 Dataset-wise Evaluation

As the first evaluation step, we compared the proposed approaches. For the German ALICE-LP dataset, we used *GBERT* and *GELECTRA* (Chan et al., 2020), established transformer encoder models pre-trained specifically for German as the basis for all encoder-based implementations.[2] For the English *ASAP* dataset, we used the recently released *ModernBERT* (Warner et al., 2024) as the base encoder.[3] The reason for this choice over more established transformer language models, such as *RoBERTa* (Liu et al., 2019a), is that *ModernBERT* achieved state-of-the-art performance on multiple established benchmarks while providing a context window of 8,192 tokens, which allows us for the reading comprehensions questions in ASAP to include the full corresponding texts in the input as part of the question spans. As LLM baselines, we use *Mistral-7b-Instruct-3.0* (Jiang et al., 2023) and *Llama-3.1-8B-Instruct* namely *Llama-3.1-8b-Instruct* (Grattafiori et al., 2024), *Qwen-3-8b* (Qwen Team, 2025), *GPT-4o* and *GPT-3.5-Turbo*.

As visible in Table 5, *GRASP* performs well for the *ALICE-LP* dataset. It significantly outperforms the other approaches, especially for the *unseen questions* dataset. This suggests that the performance of *GRASP* better translates to unseen questions than the other approaches. The results for five-shot prompting are in line with the earlier findings by Ferreira Mello et al. (2025) that zero-shot prompting generative LLMs is, as of now, a subpar operationalisation for short-answer scoring.

Table 6, on the other hand, shows that the superior performance of *GRASP* over *SBERT* and *Cross Encoders* is seemingly not achievable for the *ASAP-SAS* dataset. Here, ranking via both *SBERT* and *Cross Encoders* outperforms *GRASP* by a large margin, with our *Cross Encoders* approach placing third on the overall *ASAP-SAS* leaderboard and being the best approach that works without any form of ensembling, as used by Ormerod (2022), or data augmentation, as used by Kumar et al. (2019), as of now. Overall, the performance of the different approaches for *ASAP-SAS* shows a reversed pattern compared to *ALICE-LP*.

We hypothesised that the comparably bad perfor-

[2]These models were trained by us on a local workstation computer with a Ryzen 5900X and an Nvidia GeForce RTX 3080. Gradient Accumulation was used to realise larger batch sizes without OOM.

[3]These models were trained by us via Lambda.ai using an Nvidia GH200 unit.

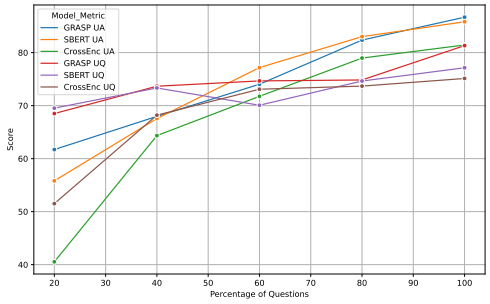

Figure 2: Learning curves depicting the influence of the percentage of questions used to train a model on downstream performance for unseen answers and unseen questions. Curves were calculated using *GBERT* as base.

mance of *GRASP* for this dataset could be due to the low number of only 10 questions. Since *GRASP* works by predicting alignment between student answers and rubrics, we hypothesised that the model might need to be trained with a diverse set of questions and rubrics to better regularise the overall task of rubric-based short answer scoring without overfitting one dataset. *ASAP-SAS* comes with many more answers per individual question, compared to *ALICE-LP*. Table 7 (Appendix) shows the question-wise performance for the *ASAP* dataset. It is visible that, for some items, *GRASP* does not fall far behind or performs better than at least one of the other approaches (3, 5, 9, 10), while this is not the case for the rest.

4.3 Influence of the Number of Training Questions on Downstream Performance

To further evaluate whether the number of different questions seen during training could be a factor for the success of *GRASP*, we conducted a learning curve study on the *ALICE-LP* dataset in which we conducted random sampling on a per-question basis, training models on 20%, 40%, 60%, 80%, and 100% of the full number of questions contained in the training set. This was conducted for the *GBERT* versions of *GRASP*, *Cross Encoders* and *SBERT*. Figure 2 depicts the corresponding learning curves. It is visible that, overall, *Cross Encoders* falls behind in all settings. For 60% and 80% of the questions, *SBERT* seems preferrable for *unseen answers*. However, in most other settings, *GRASP* is on par or better than *SBERT* for both subsets. The gap is particularly large for the 20% case.

5 Conclusion

In this paper, we introduced the task paradigm of rubric-based short answer scoring, a short answer scoring setup that, instead of relying on reference answers, as in *similarity-based scoring* (Bexte et al., 2022) or plain text classification, as in *instance-based scoring* (Bexte et al., 2022), aims at conditioning models to explicitly select a given performance level from a provided rubric by assessing a given student answer against the different performance level-wise criteria it defines. With *GRASP*, our work introduces a new operational paradigm for short answer scoring: rubric-conditioned span alignment. Though composed of known components, GRASP's architecture uniquely combines them in a way that directly mirrors human rubric use and supports generalisation to unseen rubrics.

We evaluated *GRASP* against two pairwise ranking setups implemented via *SBERT* and *Cross Encoders* (Reimers and Gurevych, 2019). This was conducted with the help of a novel German language dataset, namely *ALICE-LP*, and the established *ASAP-SAS* dataset. While *GRASP* excelled for *ALICE-LP* and outperformed the other two methods, especially with regards to transfer to unseen rubrics, it fell short for *ASAP-SAS*, where pairwise ranking via *Cross Encoder* could achieve the overall third-best reported performance, and the best out of any published approaches that do not rely on data ensembling or augmentation techniques. Overall, the results suggest that research on rubric-based short answer scoring is highly promising, and all three proposed implementations can prove valuable, depending on the exact nature of the data used, with *GRASP* particularly seeming to excel in transfer to unseen questions and rubrics.

For future work, we aim to explore the applicability of the proposed approaches to richer analytic rubrics and multilingual rubric adaptation, building toward truly generalizable scoring models. Moreover, to kickstart research on rubric-based short answer scoring in the community, we plan on making the *ALICE-LP* dataset publicly available in the context of a shared task[4]. Overall, we can conclude that rubric-based short answer scoring as a new paradigm for short-answer scoring can achieve promising results while closely mirroring human rubric use.

[4]Until then, access for purposes such as replication studies will only be granted on request, which, at the same time disqualifies the corresponding researchers for participation in the planned shared task

Limitations

Generative LLM baselines: We did not extensively evaluate generative LLMs for the problem of rubric-based short answer scoring, and, to not fully exclude them, resorted to the well-known *GTP-4o*, *GPT-3.5-Turbo* and smaller 7B and 8B models without extensive prompt tuning. This is because, going by the results of Kortemeyer (2024), Ferreira Mello et al. (2025), and Chamieh et al. (2024), even larger LLMs seem to struggle with automatic short answer scoring compared to fine-tuned transformer encoders and even older feature-based approaches, which is not just the case for short answer scoring but also for other NLP tasks (Laskar et al., 2023; Saattrup Nielsen et al., 2025). In this context, it needs to be remarked that, of course, comparing few-shot prompting to fine-tuned models is a somewhat skewed comparison. Thus, we mainly included LLM prompting to better relate our paper to the overall ongoing research context, and, since few-shot prompting requires much fewer training examples, LLMs can be seen as representative for what is possible in rubric-based short answer scoring without a need for extensive data collection.

Moreover, since LLMs have an environmental impact, and this impact is proportional to their number of parameters, on average (Bender et al., 2021), it was also a clear motivation for us to find approaches that work with fine-tuning smaller encoder-based models, which is more environmentally friendly. We still hypothesise that, with enough prompt tuning and advanced prompting techniques, especially larger generative LLMs with a number of parameters $\geq 70b$ parameters could likely also perform very well in short answer scoring. However, exploring this would have required a fundamental different focus.

Influence of variety in performance-level granularity: The two datasets we used for this study both do not possess a high variety in terms of performance-level granularity. *ALICE-LP* only has performance levels from 0 to 2, and *ASAP-SAS* comes with levels on a scale from 0 to 2 and 0 to 3. For this reason, we could not study the degree to which the performance level granularity influences the performance of *GRASP* and the other models.

Limited hyperparameter search: Due to limited resources, we only tested a small range of hyperparameters for GRASP, SBERT and Cross Encoders (3, 6, 10 epochs; 5e-6, 1e-5, 2e-5, 5e-5 learning rate; 2, 4, 8 batch size). We report the best hyperparameter combinations in the appendix.

Ethical Statement

Automatic short answer scoring is an educational NLP task. The EU AI act (European Parliament and Council of the European Union, 2024) labels AI technology (including NLP technology) in education rightfully as a high-risk application. While the individual risk depends highly on the exact context in which the corresponding technology is used and must be assessed case-by-case, mispredictions can tremendously impact learner success even in low-stakes scenarios.

For example, there is clear empirical evidence that negative feedback (and the predicted performance levels, if low, are nothing but that, if presented to a given learner) can hurt the intrinsic motivation of learners (Fong et al., 2019). If a system based on one of our presented approaches wrongfully scores correct answers as wrong, learner motivation might thus unnecessarily suffer. Even worse, when such mistakes happen in high-stakes assessments, it might negatively affect students' overall life path since, in many countries, access to university programs and jobs is highly coupled with assessment results, e.g., in the form of GPA scores. Deployment in such scenarios, therefore, requires extensive evaluation.

On the other hand, if a model is, for example, used in formative assessment and mispredicts a given wrong student answer as being correct, the corresponding student might not revise possible misconceptions present in their answer. If this happens too often throughout a given unit, students might develop misunderstandings about the content. Moreover, there is already existing research on teacher dashboards that comprehensively summarise student performance so teachers can plan interventions based on that (Karademir et al., 2024). If a non-reliable short answer scoring system powers such a dashboard, teachers might make the wrong interventions, which, in turn, could hurt student learning.

Another aspect that needs to be further assessed, which was out of the scope of this particular study, is whether the underlying models replicate undesired biases. An example of this might be a possible bias against students with dysgraphia or dyslexia. If dyslexic or dysgraphic writing is not sufficiently represented in a given training set, systems might encounter problems dealing with the same, hurt-

ing downstream predictive performance for student answers formulated by affected students.

References

Nico Andersen, Fabian Zehner, and Frank Goldhammer. 2023. Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning*, 39(3):841–854.

Lyle F. Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael J. Pan, Chris Salvador, and Yasuyo Sawaki. 2002. A reliable approach to automatic assessment of short answer free responses. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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. 2024. Strengths and weaknesses of automated scoring of free-text student answers. *Informatik Spektrum*, pages 1–9.

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, and 1 others. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

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.

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

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Laurie Ane Cutrone and Maiga Chang. 2010. Automarking: automatic assessment of open questions. In *2010 10th IEEE International Conference on Advanced Learning Technologies*, pages 143–147. IEEE.

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.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. OJ L 2024/1689, 12 July 2024.

Nigel Fernandez, Aritra Ghosh, Naiming Liu, Zichao Wang, Benoît Choffin, Richard Baraniuk, and Andrew Lan. 2022. Automated scoring for reading

comprehension via in-context bert tuning. In *International Conference on Artificial Intelligence in Education*, pages 691–697. Springer.

Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 93–103.

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.

Carlton J Fong, Erika A Patall, Ariana C Vasquez, and Sandra Stautberg. 2019. A meta-analysis of negative feedback on intrinsic motivation. *Educ. Psychol. Rev.*, 31(1):121–162.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Michael Hahn and Detmar Meurers. 2012. Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 326–336, Montréal, Canada. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Onur Karademir, Lena Borgards, Daniele Di Mitri, Sebastian Strauß, Marcus Kubsch, Markus Brobeil, Adrian Grimm, Sebastian Gombert, Nikol Rummel, Knut Neumann, and 1 others. 2024. Following the impact chain of the la cockpit: an intervention study investigating a teacher dashboard's effect on student learning. *Journal of Learning Analytics*, 11(2):215–228.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

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.

Gerd Kortemeyer. 2024. Performance of the pre-trained large language model gpt-4 on automated short answer grading. *Discover Artificial Intelligence*, 4(1).

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.

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.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Zhaohui Li, Susan Lloyd, Matthew Beckman, and Rebecca J Passonneau. 2023a. 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.

Zhaohui Li, Yajur Tomar, and Rebecca J. Passonneau. 2021. A semantic feature-wise transformation relation network for automatic short answer grading. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6030–6040, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaohui Li, Chengning Zhang, Yumi Jin, Xuesong Cang, Sadhana Puntambekar, and Rebecca J Passonneau. 2023b. Learning when to defer to humans for short answer grading. In *International Conference on Artificial Intelligence in Education*, pages 414–425. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297.

Smit Marvaniya, Swarnadeep Saha, Tejas I Dhamecha, Peter Foltz, Renuka Sindhgatta, and Bikram Sengupta. 2018. Creating scoring rubric from representative student answers for improved short answer grading. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 993–1002.

Robert J Mislevy, Russell G Almond, and Janice F Lukas. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1):i–29.

Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses.

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

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.

Qwen Team. 2025. Qwen3.

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.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Brian Riordan, Michael Flor, and Robert Pugh. 2019. How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 116–126, Florence, Italy. Association for Computational Linguistics.

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.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2025. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual NLU tasks. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 561–572, Tallinn, Estonia. University of Tartu Library.

Swarnadeep Saha, Tejas I Dhamecha, Smit Marvaniya, Renuka Sindhgatta, and Bikram Sengupta. 2018. Sentence level or token level features for automatic short answer grading?: Use both. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19*, pages 503–517. Springer.

Raheel Siddiqi and Christopher Harrison. 2008. A systematic approach to the automated marking of short-answer questions. In *2008 IEEE International Multitopic Conference*, pages 329–332. IEEE.

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.

Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using transformer-based pre-training. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*, pages 469–481. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700.

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.

Zifan Wang and Christopher Ormerod. 2024. Generative language models with retrieval augmented generation for automated short answer scoring. *arXiv preprint arXiv:2408.03811*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

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*.

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

A.1 ASAP-SAS past work baselines

Ormerod (2022) fine-tunes a range of transformer encoder models in a regular classification setup and then uses them to form an ensemble model. For this purpose, the class-wise logits of the three top-performing models, namely DeBERTa-V3-base (He et al., 2023), ELECTRA-large (Clark et al., 2020), and RoBERTa-large (Liu et al., 2019b), are fed into a multinomial logistic regression classifier, which predicts the final scores based thereon.

Kumar et al. (2019) feed Random Forests classifiers trained per individual question with an extensive and diverse set of features such as word embeddings, PoS tags, word overlap scores, keywords and sentence length. Moreover, they conduct data augmentation for each question and generate additional zero-level answers by taking highly rated answers and randomly mixing their word order.

Riordan et al. (2019) use a recurrent neural network based on bidirectional GRUs (Chung et al., 2014) whose outputs are max-pooled and fed to a linear classification head. Inputs to this network are static word embeddings concatenated with character embeddings produced by a preceding character encoder layer based on a CNN. The motivation for the latter is to account for misspellings which the static word embeddings would not be able to represent.

Ramachandran et al. (2015) use various strategies to extract regular expressions from the top-performing answers for each question in the training set. These question-wise regular expressions check for aspects such as word- or phrase-wise overlaps. Each regular expression constitutes a binary feature. Using this feature set, a separate Random Forests classifier is fit for each question.

A.2 ASAP-SAS task-wise results

Question	SBERT	CrossEnc	GRASP
1	**86.58**	85.09	76.94
2	77.76	**78.77**	73.22
3	**69.66**	65.06	66.31
4	71.65	**76.30**	64.89
5	78.86	**81.70**	80.73
6	**85.67**	85.43	77.00
7	69.68	**70.90**	52.81
8	65.27	**67.50**	61.61
9	79.81	**83.87**	77.86
10	**75.95**	75.92	74.98
Mean	76.09	**77.05**	70.64
$\text{Mean}_{\text{Fisher}}$	76.96	**77.95**	70.09

Table 7: Quadratic Weighted Kappa values achieved for the individual ASAP-SAS questions.

A.3 Best Hyperparameters

Hyperparameter	GRASP	SBERT	Cross Encoders
Learning Rate	5e-6	2e-5	2e-5
Batch Size	4	4	4
Epochs	6	3	3

Table 8: The best hyperparameter combinations for the *ALICE-LP* models.

Hyperparameter	GRASP	SBERT	Cross Encoders
Learning Rate	1e-5	2e-5	2e-5
Batch Size	8	8	8
Epochs	10	3	3

Table 9: The best hyperparameter combinations for the *ASAP-SAS* models.

A.4 Prompt for the generative LLMs

The following task prompt was used to prompt the generative LLM baselines. For *ALICE-LP*, we used a literal German translation of this prompt to match the language of the dataset.

```
Your task is to assess the answer to the
question using the rubric to determine
the right score. Compare the answer
with the criteria provided in the rubric
for each score and assign the most
appropriate score. Always end your
response with the appropriate score
from the rubric, e.g., "Score: 0",
"Score: 1", or "Score: 2".

Question: "{question}"
Rubric: "{rubric}"
Answer: "{answer}"
Score: {score}
```

Few-shot learning is implemented by repeating this prompting scheme for each datapoint, wrapped in the corresponding special tokens that signify user and assistant parts of the prompt. Moreover, as a system prompt, the model was given the following:

```
You are a teacher who conscientiously
assesses the work of your students.
```