Rubrics as Semantic Subspaces: A Unified Approach to Rubric-based Constructed Response Scoring across Short Answers and Essays

Sebastian Gombert[1], **Sonja Hahn**[1], **Nico Andersen**[1], **Leon Camus**[1], **Zhifan Sun**[1], **Ngoc Nhu Hao Nguyen**[1], **Fabian Zehner**[1,2], **Longwei Cong**[1], **Alexander Mehler**[4] **Hendrik Drachsler**[1,3,4]

[1]DIPF | Leibniz-Institute for Research and Information in Education
[2]Centre for International Student Assessment (ZIB)
[3]Studiumdigitale & [4]Computer Science Department, Goethe University Frankfurt

Abstract

Rubrics are the primary reference for manual scoring of constructed responses, and there is growing interest in their use in automated scoring methodologies. In this work, we propose Aspect-Grounded Rubric–Answer Alignment (AGRAA), a rubric-based end-to-end scoring framework that models rubric descriptors as latent aspect spaces. Concretely, rubric descriptors are represented as low-dimensional subspaces derived from contextualised transformer embeddings, and student responses are scored according to how strongly their representations align with these rubric-induced spaces relative to the residual space outside them. This formulation provides a geometrically grounded interpretation of rubric-based scoring while enabling end-to-end training with standard transformer encoders. We introduce three distinct architectural variants and evaluate them on multiple short-answer and essay scoring datasets. Across these tasks, AGRAA achieves predictive performance highly competitive with strong neural and feature-based baselines. In addition, the framework yields interpretable intermediate representations that expose which rubric-defined aspects contribute to scoring decisions, enabling decision-aligned explanations grounded in rubric descriptors.

1 Introduction

Rubric-based scoring plays a central role in assessing constructed responses in educational settings. Rubrics describe performance using qualitative criteria that enable fine-grained evaluation of responses (Reddy and Andrade, 2010). As automated constructed-response assessment becomes increasingly important in digital learning environments (Mello et al., 2024) and large-scale testing (Zehner et al., 2025), there is growing interest in methods that can automatically score responses while leveraging the criteria defined in the rubrics (Sonkar et al., 2024; Wang et al., 2019; Gombert et al., 2026; Cong et al., 2026; Sreedevi et al., 2026; Eltanbouly et al., 2025).

Across domains and response types, rubrics typically describe performance in terms of aspects such as conceptual correctness, completeness, use of evidence, or argumentative quality (Reddy and Andrade, 2010). These aspects are usually distributed across the response rather than localised to individual words or spans. For example, a science rubric might require identifying oxidation, explaining oxygen exposure, and linking the process to the observed color change in an apple. Scoring therefore amounts to estimating how strongly a response expresses the aspects associated with each rubric level.

In this work, we frame rubric-based scoring as a form of *neural subspace classification* (Fukunaga, 2013; Fukui, 2021) on top of contextual embeddings produced by transformer encoder models. Each rubric level defines a low-dimensional semantic subspace derived from the dominant directions of variation in its textual representation. These directions can be interpreted as latent aspects of the rubric. Student responses are projected into these rubric-induced spaces, and scoring is implemented by measuring how much of the response representation lies within each rubric subspace relative to the residual space outside it. The summarised contributions of this paper are:

- We introduce **AGRAA**, a rubric-based representation learning framework that aligns student responses with rubric descriptors through subspace-based similarity, enabling rubric-aware scoring decisions that expands upon earlier ideas introduced by Gombert et al. (2026).
- We propose three architectural variants: a *bi-encoder* formulation (Bi-AGRAA), a *level-wise cross-encoder* formulation (Cross-AGRAA), and a *joint cross-encoder* formulation (Joint-Cross-AGRAA) that allows direct

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 624–634

interaction between rubric levels during representation learning.

- Through experiments on multiple short-answer and essay scoring datasets, we demonstrate that rubric-based subspace alignment achieves **competitive predictive performance** while providing transparent, rubric-aligned explanations of model decisions.

- We show that the approach yields **interpretable intermediate representations** that reveal which aspects within rubric descriptors contribute to a prediction.

2 Background

2.1 Automated Assessment of Constructed Responses

The automated assessment of constructed responses is a long-established research branch in educational natural language processing (Bai and Stede, 2022; Burrows et al., 2014; Bexte et al., 2024). In prior research, two genres of constructed responses, namely essays and short answers, were the main research focus, with distinct research trajectories for both genres (Bai and Stede, 2022). While short answer scoring mainly focuses on the content of given answers ("content scoring") (Bexte et al., 2024; Ziai et al., 2012), essay scoring is mainly focused on writing-level traits such as coherence, grammar, argumentation, or style (Bai and Stede, 2022), although there is some work that instead deals with essay content scoring (Gombert et al., 2024).

2.2 Rubrics in Automated Short Answer Scoring (ASAS)

Scoring rubrics have previously been explored as inputs to automatic short-answer scoring systems. Wang et al. (2019) proposed a BiLSTM-based architecture that computes word-level alignment scores between tokens in a student response and tokens in rubric descriptors. They showed that incorporating rubric information improves performance compared to a plain BiLSTM-based scoring model introduced by Riordan et al. (2017), particularly in low-resource settings. These results provide early evidence that explicit modeling of rubric information can benefit automatic short-answer scoring.

Gombert et al. (2026) proposed two architectures for rubric-based short answer scoring. Both architectures use attention to calculate alignment scores between responses and rubric descriptors using embeddings from transformer encoder models. The performance level whose corresponding rubric has the highest alignment to a given response is assigned as the final score. The variants differ in how alignment is calculated, on top of mean-pooled span embeddings or between the embeddings of individual rubric descriptor and response tokens.

Cong et al. (2026) and Wei et al. (2025) evaluated the utility of rubrics for short answer scoring when using LLM in-context learning. In both cases, providing the LLM with rubrics significantly improved performance compared to a baseline without rubrics and was more beneficial than providing simple few-shot examples.

2.3 Rubrics in Automated Essay Scoring

Sreedevi et al. (2026) proposed a rubric-aware neural architecture for automated essay scoring that explicitly incorporates analytic rubric criteria into the modeling process. Essays are encoded using contextual embeddings augmented with discourse information. These discourse-aware essay representations are then connected to rubric traits through trait-specific attention mechanisms, enabling the model to learn rubric-conditioned representations of essay quality. The model jointly predicts both holistic essay scores and rubric-level trait scores, with the rubric acting as a structured supervisory signal that guides the decomposition of essay quality into interpretable evaluation dimensions.

Eltanbouly et al. (2025) implemented an essay scoring framework that uses LLMs to extract features from given rubrics that are then used in a feature-based statistical classifier. Concretely, they prompt LLMs to transform given rubrics into binary yes-no questions that ask whether a given rubric-defined aspect is fulfilled by the essay or not. In the next step, the LLM is prompted to answer these individual questions. The resulting binary features are used as input features for a logistic regression classifier.

3 Aspect-Grounded Rubric-Answer Alignment (AGRAA)

Rubrics typically describe the quality of the answer in terms of qualitative criteria such as conceptual correctness, completeness, use of evidence, or quality of the explanation (Reddy and Andrade, 2010; Popham, 1997). Importantly, these descriptions rarely correspond to a single lexical cue or an iso-

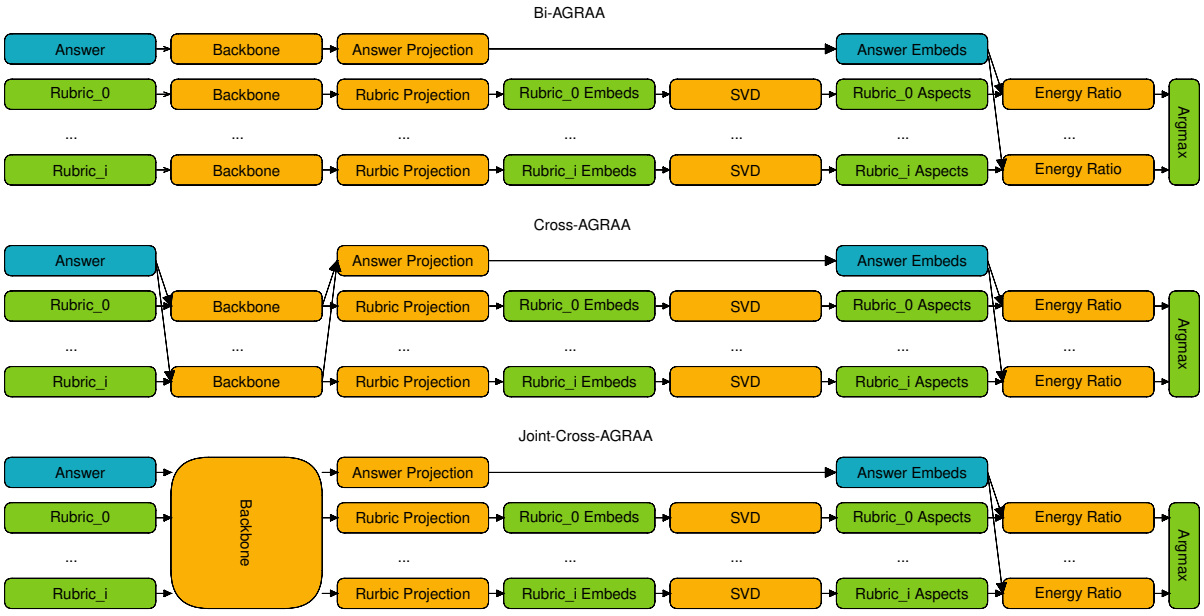

Figure 1: A schematic depiction of both architectures. SVD = Singular Value Decomposition. Questions are included as context in each forward pass.

lated fact. Instead, rubric levels implicitly refer to *aspects* that characterise responses at different quality levels (Popham, 1997). For example, a "correct" rubric level in a science task may simultaneously require the identification of a process, explanation of a mechanism, and appropriate use of domain terminology. These aspects are distributed across the response (Popham, 1997) and can be realised through a wide range of linguistic expressions (Kreidler, 1998).

Student responses also do not express rubric-relevant aspects in a single location. Rather, aspects are typically distributed throughout the text and expressed with varying degrees of completeness and clarity. We therefore can interpret both rubric levels and student responses as *configurations of aspects*. Each rubric level defines a set of aspects expected for that level, while each response expresses these aspects to varying degrees. The central task of rubric-based scoring can thus be framed as estimating how strongly a response expresses the aspects associated with each rubric level. Therefore, our approach is based on three core assumptions:

1. Each rubric level descriptor defines a set of latent aspects.

2. The responses express these aspects with varying strength.

3. Scoring corresponds to measuring the alignment between the aspects defined in the rubric and their materialisation in a given response.

Based on these assumptions, we implemented *AGRAA*, an architectural framework for end-to-end scoring of constructed responses grounded in subspace classification methodology (Fukui, 2021).

3.1 Encoder and Asymmetric Projection

Given a question/item q and a response a, we aim to predict the correct performance level $0, \ldots, i$ for answer a based on the qualitative criteria described in rubric R consisting of performance level-wise descriptors $r_0, \ldots, r_i$. For this purpose, we consider three architectural variants that differ in how contextual interactions between the response and the rubric are modeled. Figure 1 depicts these variants schematically.

Bi-AGRAA. The first variant is a *bi-encoder* setup (*Bi-AGRAA*), where the response and each rubric level are encoded independently. In this case, the response representation is shared across all rubric levels. The question is added as contextual input at each forward pass but masked during all downstream processing steps.

Cross-AGRAA The second variant is a *level-wise cross-encoder* setup (*Cross-AGRAA*), where the encoder jointly processes the response together with a single rubric level at a time. This produces response representations that are conditioned on the

rubric level under consideration. The question is added as contextual input at each forward pass but masked during all downstream processing steps.

Joint-Cross-AGRAA The third variant is a *joint cross-encoder* setup (*Joint-Cross-AGRAA*), where the question, response, and all rubric levels are encoded together in a single forward pass. This allows for direct contextual interaction between rubric levels while producing a single globally contextualised response representation. The question is added as contextual input at each forward pass, but is masked for all downstream processing steps. This is the exact setup used by Gombert et al. (2026) for their architectures.

In all variants, functional tokens such as CLS or SEP are masked for all following computations to prevent the models from learning uninterpretable shortcuts involving these tokens. Moreover, all tokens of q are masked for all steps after encoding. Each of the three variants yields contextualised embeddings for response and rubric tokens:

$$H_a \in \mathbb{R}^{T_a \times d}, \qquad H_{r_i} \in \mathbb{R}^{T_{r_i} \times d} \tag{1}$$

In these, d denotes the encoder hidden size. To account for the different functional roles of rubric and response text (the former describes in which aspects a correct response materialises, while the latter is a concrete materialisation), we apply asymmetric linear projections into a shared comparison space of dimension l:

$$X_a = H_a W_a + B_a, \qquad X_a \in \mathbb{R}^{n_a \times l} \tag{2}$$

$$X_{r_i} = H_{r_i} W_r + B_r, \qquad X_{r_i} \in \mathbb{R}^{n_{r_i} \times l} \tag{3}$$

Separate projection layers allow rubric tokens to define aspect directions while response tokens are evaluated relative to these directions.

3.2 Rubric-induced Aspect Spaces

On top of the three underlying encoding approaches, we use the same specialised head, which can be viewed as a neural subspace classifier (Fukunaga, 2013) in which each rubric level descriptor defines a semantic subspace of one or more latent aspects. In this perspective, scoring corresponds to a form of nearest-subspace classification where responses are assigned to the rubric level whose semantic subspace best explains their representation. We estimate these aspects by constructing a low-dimensional subspace from the rubric token embeddings. Given projected rubric token embeddings X_{r_i}, we compute a singular value decomposition:

$$X_{r_i} = U_i \Sigma_i V_i^\top \tag{4}$$

The right singular vectors define orthogonal directions in the projection space capturing dominant semantic variation within the rubric text. We interpret these directions as latent aspects. To determine the number of aspects a given rubric descriptor contains, which of course differs from descriptor to descriptor, we retain the smallest number of singular directions explaining a fixed proportion of variance τ (e.g., 0.9), capped by a maximum number of aspects $k_{\max}$. Both need to be set as hyperparameters depending on the granularity of the rubrics at hand:

$$k_i = \min\Big(\{k_{\max}\} \cup \{k : \tfrac{\sum_{j=1}^{k} \sigma_{ij}^2}{\sum_j \sigma_{ij}^2} \geq \tau\}\Big) \tag{5}$$

This yields an orthonormal basis $Q_i \in \mathbb{R}^{r \times k_i}$ representing the aspect space associated with rubric level i, together with singular values σ_{ij}.

3.3 Response-to-Aspect Alignment

To evaluate how strongly a response expresses each rubric aspect, we project response token embeddings into the rubric-induced subspace:

$$A_i = X_a Q_i \in \mathbb{R}^{n_a \times k_i} \tag{6}$$

Each column of A_i represents the strength with which the response tokens align with a particular rubric aspect. To quantify how well a response can be expressed within the rubric-induced aspect space, we compute the proportion of response variance explained by this subspace. Let

$$E_{\parallel,i} = \frac{1}{n_a} \|A_i\|_F^2 \tag{7}$$

denote the energy of the response within the rubric-induced subspace, and

$$E_{\text{tot}} = \frac{1}{n_a} \|X_i\|_F^2 \tag{8}$$

The total response energy in projection space. The residual energy outside the rubric subspace is:

$$E_{\perp,i} = E_{\text{tot}} - E_{\parallel,i}. \tag{9}$$

We then define the alignment between a response and rubric level i as the log-ratio between the energy explained by the rubric-induced subspace and the residual energy outside the subspace, a form closely related to the classical subspace classification criteria in statistical pattern recognition (Fukunaga, 2013):

$$s_i = \log \frac{E_{\parallel,i} + \varepsilon}{E_{\perp,i} + \varepsilon}. \tag{10}$$

3.4 Classification and Training

The predicted scores are obtained by an argmax in all $s_0, \ldots, s_i$:

$$\hat{y} = \arg\max_i s_i. \tag{11}$$

The model is trained end-to-end using softmax cross-entropy loss over these alignment scores.

Aspect distribution regularisation. While the above objective optimises predictive performance, early pre-experiments showed that the resulting models often collapsed onto one dominant "master" aspect that largely determined correctness, while the remaining directions acted as residual aspects without a clear function. This constitutes a learned shortcut. However, as stated earlier, rubrics typically describe performance across multiple aspects. To encourage the model to better reflect this multi-aspect structure, we introduce an additional regularisation term to the loss function that promotes a more distributed use of rubric-based semantic directions.

To obtain a direction-level summary of how strongly each aspect contributes to the alignment with a rubric unit, we first compute pooled per-direction alignment values:

$$e_{i,j} = \sqrt{\frac{1}{\alpha} \ln \sum_{t=1}^{n_a} \exp\big(\alpha\, (A_i)_{t,j}^2\big)}. \tag{12}$$

These values summarise the strongest token-level evidence for each semantic direction. We then derive a normalised distribution over aspect contributions

$$p_k = \frac{e_{i,k}}{\sum_{m=1}^{k_i} e_{i,m}}, \tag{13}$$

and compute its entropy

$$\mathcal{H}(p) = -\sum_{k=1}^{k_i} p_k \log p_k. \tag{14}$$

In practice, this entropy is normalised with respect to the number of retained directions and is used as a soft constraint: the model is penalised only when the aspect distribution becomes too concentrated, that is, when its entropy falls below a predefined target range. This encourages the model to make use of multiple rubric-defined semantic directions without forcing all directions to contribute equally. Because regularisation operates only on the dynamically retained directions for a given rubric unit through variance-based subspace construction, it promotes a balanced use of the available semantic directions without forcing the model to use the maximum possible number of aspects in every case. The final loss function is defined as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{ent}}, \tag{15}$$

where $\mathcal{L}_{\text{ent}}$ denotes the entropy-based penalty and λ controls its strength.

3.5 Interpretability

To obtain decision-aligned explanations, we decompose the final prediction into contributions of individual aspects. Let $e_{i,j}$ denote the pooled response energy along the aspect j of rubric level i (as defined in equation 12), and $w_{i,j}$ the corresponding normalised singular-value weight. Within each rubric level, aspect contributions are computed as:

$$\tilde{c}_{i,j} = \frac{w_{i,j}\, e_{i,j}^2}{\sum_m^{k_i} w_{i,m}\, e_{i,m}^2}. \tag{16}$$

To align the explanation with the model's cross-level decision, these quantities are weighted by softmax-normalised level scores:

$$c_{i,j} = \text{softmax}(s)_i \cdot \tilde{c}_{i,j}. \tag{17}$$

The values $c_{i,j}$ sum to one across all levels and aspects and therefore directly quantify how much each rubric-defined semantic direction contributed to the final classification. The resulting visualisations make explicit which rubric aspects drove the decision and which rubric tokens characterise these aspects (see Figure 2). This yields a decision-aligned explanation in terms of rubric semantics. Because explanations are derived directly from the same projected response representations used in the scoring function, the resulting aspect contributions reflect the internal evidence underlying the model's decision rather than a post-hoc approximation.

3.6 Applicability across response types

Given a long-context encoder is used as backbone, the architecture can support constructed response types of varying length. Short answers and essays, as the most important types of written constructed responses, differ primarily in the number, focus, and complexity of aspects expressed, as well as in overall length. This can be accommodated by adjusting the maximum number of retained aspect directions $k_{\max}$.

Approach	UA P	UA R	UA F1	UA QWK	UQ P	UQ R	UQ F1	UQ QWK
ModernGBERT-1B Joint-Cross-AGRAA	**.760**	**.740**	**.743**	.748	.620	.577	.581	.539
ModernGBERT-1B Cross-AGRAA	.745	.734	.736	.733	.594	.584	.586	.521
ModernGBERT-1B Bi-AGRAA	.726	.711	.714	.705	.567	.552	.558	.475
ModernGBERT-1B Classifier (Q+A+Ru)	.732	.716	.719	.713	.579	.553	.553	.466
ModernGBERT-1B Classifier (Q+A)	.707	.704	.705	.692	.533	.529	.522	.404
GBERT-large ToLeGRAA (Gombert et al., 2026)	.748	.739	.742	**.751**	.620	.595	.600	.569
Claude Sonnet 4.5 5-shot (Gombert et al., 2026)	.680	.679	.679	.665	**.646**	**.640**	**.641**	**.603**

Table 1: Results achieved for the ALICE-LP dataset. We follow the evaluation protocol established in Gombert et al. (2026) and report dataset-wide metrics for the unseen answers and unseen questions evaluation sets. P = Weighted Precision. R = Weighted Recall. F1 = Weighted F1. QWK = Quadratic Weighted Kappa. UA = Unseen Answers subset. UQ = Unseen Questions subset.

4 Evaluation

Following the definitions from Murdoch et al. (2019), we evaluate the models for both *predictive* and *descriptive* accuracy. While predictive accuracy refers to the predictive performance of models, descriptive accuracy concerns whether these models make predictions for plausible reasons, i.e., whether plausible aspects are extracted from the rubrics and whether the calculated alignment matches what is stated in a given input answer. Our evaluation, therefore, addresses two concrete research questions:

RQ1. What is the *predictive accuracy* of the three *AGRAA* variants?

RQ2. Are the models faithful in terms of *descriptive accuracy* in regard to the content of answers and rubrics?

For the first research question, we measure prediction quality, while for the second research question, we qualitatively analyzed 64 randomly selected examples from the unseen-answers evaluation set of the *ALICE-LP* dataset (see next section) for each model variant (Bi-AGRAA, Cross-AGRAA, Joint-Cross-AGRAA).

4.1 Datasets

For our experiments, we used two datasets for short-answer scoring and one for holistic essay scoring, namely *ALICE-LP*, *ASAP-SAS*, and *PERSUADE 2.0*. In every evaluation, the models are trained on the whole respective dataset.

ALICE-LP 1.0. This short answer scoring dataset was first introduced by Gombert et al. (2026) and includes 13,075 German short answers to 118 questions from four different STEM domains collected in German middle and high schools. All answers are rated on a three-point scale following question-specific rubrics. It is divided into a test and two evaluation sets. One of these contains answers to questions seen during training (*unseen answers*), while the other contains answers to questions unseen during training (*unseen questions*).

ASAP-SAS. This short-answer scoring dataset was first introduced in the context of a 2012 Kaggle competition (Barbara et al., 2012) and has since grown into a widely used benchmark for automated short-answer scoring. It comprises 22,067 English answers to a total of ten different questions with question-specific rubrics, including both STEM and reading comprehension questions, collected at US-American schools.

PERSUADE 2.0. This essay scoring dataset was first introduced by Crossley et al. (2024). It comprises 25,996 argumentative essays written in English for 15 different prompts. It includes holistic scores on a 6-point scale, as well as analytic scores for multiple argumentation-related criteria on a 3-point scale. In this work, we focus on predicting the holistic scores. The dataset contains source-based and independent writing prompts, both of which are scored using slightly different rubrics.

4.2 Backbone Models, Hyperparameters and Baselines

As backbone models, we used *ModernBERT* (Warner et al., 2025) variants. *ModernBERT* is a modified variant of the original *BERT* architecture that supports long-context processing with the help of RoPE embeddings (Su et al., 2024). For the German *ALICE-LP* set, we use *ModernGBERT-1B* (Wunderle et al., 2025), a German-specific ModernBERT (Warner et al., 2025) variant, while for the English *ASAP-SAS* and *PERSUADE 2.0*, we use the regular *ModernBERT-large* (Warner et al., 2025). As baselines, we employ two classifier vari-

Approach	1	2	3	4	5	6	7	8	9	10	QWK_m	QWK_F
ModernBERT-large Joint-Cross-AGRAA	.871	**.883**	.709	**.760**	.796	**.887**	.714	.667	.830	.782	.790	.802
ModernBERT-large Cross-AGRAA	.874	.878	.715	.717	.810	.883	.732	.664	.846	.749	.786	.800
ModernBERT-large Bi-AGRAA	**.885**	.863	.706	.748	.837	.835	.706	.679	.842	.740	.784	.795
ModernBERT-large Classifier (Q+A)	.857	.831	.729	.707	.802	.806	.716	.658	.825	.759	.769	.776
ModernBERT-large Classifier (Q+A+Ru)	.859	.816	.654	.730	.791	.803	.732	.641	.829	.783	.764	.772
Ensemble of fine-tuned MLMs (Ormerod, 2022)	.882	.868	.722	.750	.813	.822	.734	.702	**.865**	.779	**.796**	**.803**
Random Forests + div. feats. (Kumar et al., 2019)	.872	.824	**.745**	.743	**.845**	.858	.725	.624	.843	**.832**	.791	.802
GTE-EN-MLM Classifier (Q+A+R) (Gombert et al., 2026)	.868	.848	.715	.738	.790	.827	.681	**.718**	.844	.773	.780	.788
ModernBERT-large GRAASP (Gombert et al., 2026)	.876	.852	.724	.727	.803	.871	.702	.672	.802	.749	.779	.788
GRU (Riordan et al., 2019)	.830	.791	.662	.731	.844	.861	**.736**	.664	.809	.777	.771	.779
ModernBERT-large Classifier (Q+A+Ru) (Gombert et al., 2026)	.856	.836	.677	.703	.768	.831	.699	.675	.817	.748	.761	.775

Table 2: Results achieved for the ASAP-SAS dataset. We follow the evaluation protocol used by Riordan et al. (2019), Kumar et al. (2019), Gombert et al. (2026), and Ormerod (2022) to report question-wise Quadratic Weighted Kappa scores, which are then averaged using the arithmetic (QWK_m) as well as the Fisher-weighted (QWK_F) mean.

ants of these models. For the first variant, a given classifier receives a response and the corresponding question/item prompt, and for the second variant, the classifier is also given the full rubrics. For *ALICE-LP* and *ASAP-SAS*, we set $k_{\max} = 3$, and for *PERSUADE 2.0*, we set $k_{\max} = 12$, based on a manual inspection of the number of aspects contained in the corresponding rubric descriptors. We set $\tau = 0.9$, $\lambda = 0.2$, and $\alpha = 5.0$ in all cases. Training was run for 12 epochs with a learning rate of 1e-5, a weight decay of 0.1, gradient clipping set to 1.0, a batch size of 16, and early stopping active using *AdamW* as optimiser. We report results for the best checkpoint per run.

5 Results

5.1 Predictive Accuracy

Across all evaluated datasets, the proposed AGRAA variants achieve competitive predictive performance. When compared against conventional classifier architectures built on the same underlying backbone models, several AGRAA variants consistently match or outperform these baselines. On the ALICE-LP dataset, *Joint-Cross-AGRAA* reaches performance comparable to the previously reported ToLeGRAA model on the unseen answers setting, while all AGRAA variants outperform both rubric-aware and rubric-free classifier baselines implemented with the same encoder. On the ASAP-SAS dataset, the AGRAA variants likewise improve over the corresponding classifier baselines and achieve results competitive with strong published neural and feature-based systems, with *Joint-Cross-AGRAA* and *Cross-AGRAA* closely approaching the best reported aggregate QWK scores. On the essay scoring dataset, the proposed models perform on par with or slightly above strong transformer-based baselines.

Approach	P	R	F1	QWK
ModernBERT-large Cross-AGRAA	**.694**	**.693**	**.691**	**.878**
ModernBERT-large Joint-Cross-AGRAA	.690	.689	.688	.873
ModernBERT-large Bi-AGRAA	.672	.668	.666	.866
ModernBERT-large Class. (Q+A+Ru)	.681	.679	.675	.869
ModernBERT-large Class. (Q+A)	.668	.667	.664	.864
DeBERTa-v3 Class. (Ravindran and Choi, 2025)	-	-	-	.870
ModernBERT Class. (Ormerod, 2025)	-	-	-	.868

Table 3: Results achieved for the PERSUADE 2.0 dataset. We report weighted precision (P), weighted recall (R), weighted F1 score, and quadratic weighted Kappa.

Taken together, these results indicate that the rubric-based subspace formulation underlying AGRAA can match, and, in several cases, exceed the predictive performance of conventional classifier formulations using the same encoder backbones.

5.2 Descriptive Accuracy

Overall, the Bi-AGRAA and Joint-Cross-AGRAA variants frequently decomposed rubric representations into several semantically meaningful aspects that corresponded to distinct rubric criteria or conceptual components.

Across most inspected visualisations, the extracted aspects highlighted meaningful content words from the rubric descriptors as well as modifiers associated with them, such as determiners or descriptive adjectives (e.g., *die*, *der*). This suggests that the learned semantic directions are anchored in linguistically and semantically plausible parts of the rubric descriptors.

Figure 2 illustrates an explanation produced by the *Joint-Cross-AGRAA* model. Within the predicted rubric level, three aspects that characterise the expected reasoning are visible. The

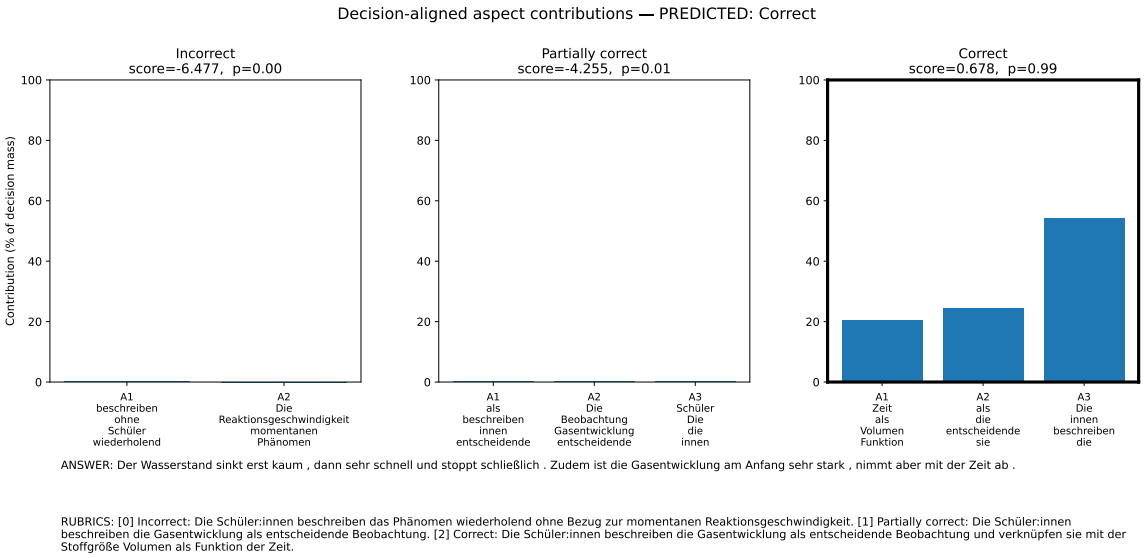

Figure 2: This figure shows an example of the aspect contributions with a correct answer from the ALICE-LP dataset the system rightfully judged as correct. Each bar represents the $c_{i,j}$ (see Equation 17) for a given score s_i (see Equation 10) and a given aspect j for a Cross-AGRAA *ModernGBERT*-1B model trained with $k_{max} = 3$. Since the model is highly confident, there are only bars visible in the correct cell. For each aspect, we show the four most characteristic words from the respective rubric.

most influential aspect is associated with the token *beschreiben*, reflecting the rubric requirement that the response should provide a description. A second aspect is characterised by tokens such as *Zeit*, *Volumen*, and *Funktion*, corresponding to the rubric requirement that the gas volume should be interpreted as a function of time. A third aspect reflects observational terminology such as the adjective *entscheidende* in *Beobachtung*, capturing the emphasis on identifying the gas development as the key phenomenon. The student response indeed describes how the gas development changes over time ("*Gasentwicklung . . . nimmt aber mit der Zeit ab*"), which aligns closely with these rubric-defined aspects, suggesting that the extracted directions correspond to semantically meaningful components of the rubric. Across all the inspected examples, we could observe similar patterns of align

6 Discussion

The experimental results suggest that rubric-based scoring can be effectively framed as a subspace alignment problem between responses and rubric descriptors. Across three datasets covering short answer- and essay scoring, the proposed **AGRAA** variants achieved predictive performance stronger than plain transformer classifier baselines in the majority of cases. The strongest among these across all datasets is the **Joint-Cross-AGRAA** variant, indicating that cross-attention between rubric descriptors as well as between rubric descriptors and answers is needed to best differentiate between performance levels. Overall, we can conclude that neural subspace classification as employed for **AGRAA** can be seen as beneficial for predictive performance in rubric-based constructed response scoring.

Moreover, the subspace formulation yields interpretable intermediate representations that expose how rubric-defined aspects contribute to scoring decisions. Qualitative analysis indicated that these aspects frequently correspond to meaningful rubric aspects, suggesting that the learned semantic directions capture pedagogically relevant structure present in the rubrics. However, it needs to be confirmed whether this pattern holds up across varied models and datasets, which needs to be the objective of future work.

Limitations

Several limitations remain. First, the interpretability analysis in this work is primarily qualitative and future work could investigate quantitative measures of explanation quality. Second, the number of retained semantic directions is controlled through global hyperparameters, which may not always perfectly reflect the conceptual granularity of a given rubric. We did not conduct an extensive hyperparameter search. A small-scale stability analysis (see

Appendix) revealed that the predictive performance stays stable under varied hyperparameters, but may experience drops for certain configurations. Finally, the approach assumes that rubric descriptors sufficiently capture the semantic structure of the scoring task, which may not hold for poorly specified rubrics. Therefore, future work would need to test the relationship between variables such as linguistic and conceptual uncertainty within rubrics and the resulting predictive performance.

Ethical Statement

Automatic constructed response 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 incorrect interventions, which, in turn, could harm 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, hurting downstream predictive performance for student answers formulated by affected students.

References

Xiaoyu Bai and Manfred Stede. 2022. 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.

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

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2024. Strengths and weaknesses of automated scoring of free-text student answers. *Informatik Spektrum*, 47(3–4):78–86.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2014. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Longwei Cong, Leon Hammerla, Sonja Hahn, Sebastian Gombert, Hendrik Drachsler, and Ulf Kröhne. 2026. Automatic short answer grading with llms: From memorization to reasoning. In *Proceedings of the 16th International Learning Analytics and Knowledge Conference (LAK '26)*, New York, NY, USA. ACM.

Scott Andrew Crossley, Yu Tian, Perpetual Baffour, A. Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.

Sohaila Eltanbouly, Salam Albatarni, and Tamer Elsayed. 2025. TRATES: Trait-specific rubric-assisted cross-prompt essay scoring. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20528–20543, Vienna, Austria. 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.

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

Kazuhiro Fukui. 2021. Subspace methods. In *Computer vision: a reference guide*, pages 1221–1224. Springer.

Keinosuke Fukunaga. 2013. *Introduction to statistical pattern recognition*. Elsevier.

Sebastian Gombert, Aron Fink, Tornike Giorgashvili, Ioana Jivet, Daniele Di Mitri, Jane Yau, Andreas Frey, and Hendrik Drachsler. 2024. From the automated assessment of student essay content to highly informative feedback: a case study. *International Journal of Artificial Intelligence in Education*, 34(4):1378–1416.

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

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*, 11:215–228.

Charles Kreidler. 1998. *Introducing English semantics*. Routledge.

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.

Rafael Ferreira Mello, Elyda Freitas, Luciano Cabral, Filipe Dwan Pereira, Luiz Rodrigues, Mladen Rakovic, Jackson Raniel, and Dragan Gaševic. 2024. Words of wisdom: A journey through the realm of natural language processing for learning analytics–a systematic literature review. *Journal of Learning Analytics*, 11(3):82–105.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

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

Christopher Ormerod. 2025. Automated essay scoring incorporating annotations from automated feedback systems. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 9–18, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

W James Popham. 1997. What's wrong-and what's right-with rubrics. *Educational Leadership*, 55:72–75.

Renjith Ravindran and Ikkyu Choi. 2025. Investigating adversarial robustness in LLM-based AES. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 86–91, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

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

Brian Riordan, 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.

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.

N. Sreedevi, M. Madhusudhan Rao, Sridevi Dasam, Roopa Traisa, Jasgurpreet Singh Chohan, V. Saranya, and Ahmed I. Taloba. 2026. Rubric-relational discourse modeling with counterfactual explainability for multi-trait automated essay scoring. *International Journal of Advanced Computer Science and Applications*, 17(2).

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568(127063):127063.

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

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

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

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)*, pages 660–671, Vienna, Austria. Association for Computational Linguistics.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 190–200, Montréal, Canada. Association for Computational Linguistics.

A Appendix

Sensitivity analysis of $k_{\max}$ and λ. To examine the sensitivity of the model to the two central structural hyperparameters, we conducted a targeted ablation for *Joint-Cross-AGRAA* on the ALICE-LP unseen-answers setting. We varied the cap on the number of retained rubric directions $k_{\max} \in \{3, 4, 5, 6\}$ and the entropy regularisation weight $\lambda \in \{0.0, 0.2, 0.6\}$ while keeping all other hyperparameters fixed. Due to computational constraints, the analysis was performed with a single random seed and should therefore be interpreted as a compact sensitivity check rather than an exhaustive hyperparameter search. Across most configurations, performance remains within a relatively narrow range, with some outliers present. This suggests that k_{max} and λ can influence performance and should therefore not be set randomly, but overall, the model is able to handle varying configurations of these hyperparameters.

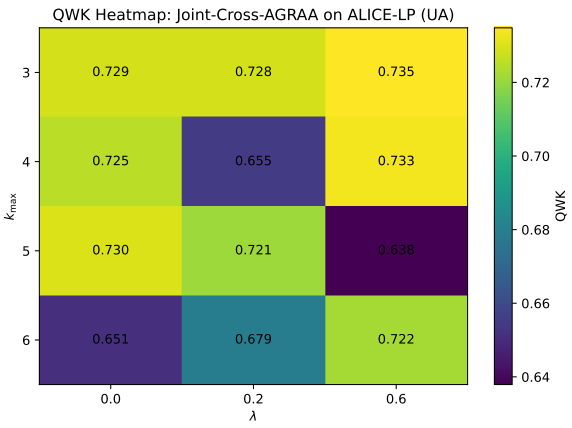

Figure 3: QWK results for the different k_max and λ combinations.