ASLAN at BEA 2026 Shared Task 2: Voting Across Scoring Paradigms

Marie Bexte[1], Yuning Ding[2,3], Josef Ruppenhofer[1], Nils-Jonathan Schaller[2,3], Daniel Mora Melanchthon[2,3], Torsten Zesch[1], Andrea Horbach[2,3],
[1]CATALPA, FernUniversität in Hagen, Germany,
[2]Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[3]Kiel University, Germany

Abstract

This paper describes the ASLAN system contribution to the BEA 2026 Shared Task on rubric-based short answer scoring for German (Gombert et al., 2026). We investigate three complementary modeling paradigms: similarity-based scoring, instance-based classification, and rubric-prompted large language models (LLMs). For the *unseen answers* track, where test answers belong to prompts observed during training, we compare question-specific and generic scoring models as well as ensemble variants. For the *unseen questions* track, where models must generalize to previously unseen prompts, we primarily rely on zero-shot LLM-based scoring using the scoring rubrics. Our experiments show that similarity-based models outperform instance-based models and LLM-based models in the *unseen answers* setting. In addition, we find that ensemble methods improve robustness over individual models.

1 Introduction

Automated Short Answer Scoring (SAS) is the task of automatically judging the conceptual correctness of learner answers in an educational context (Bai and Stede, 2023; Bexte et al., 2024).

While traditional SAS systems often rely on the comparison with a reference answer or learning properties of a correct or incorrect answer from training instances for a specific question (Bexte et al., 2023, 2022), this Shared Task aims at leveraging the scoring rubrics provided for individual answers as the basis for determining the correctness of an answer - a so far underexplored variant of SAS, that is especially relevant in those cases where no training data for a specific question is provided. The shared task consists of two tracks: *unseen answers* where the test data contains new answers to questions already seen during training and *unseen questions* with completely new questions not covered in the training material. For both tracks, there is a 3-way condition and a 2-way condition. In the 3-way setup, answers are labeled *correct*, *partially correct* or *incorrect*. The 2-way condition subsumes the *partially correct* and *incorrect* answers in a joint group of *incorrect* answers.

Our system compares several approaches: In the *unseen answers* track, we use similarity-based approaches, relying on the semantic similarity to a reference answer, instance-based approaches that learn properties of correct and incorrect answers to the same question, and LLM-based approaches leveraging generative AI and using the scoring rubric in the prompt. Only the third approach is directly applicable to the *unseen questions* track data. We further experiment with ensembles of classifiers, performing majority voting to determine the final label.

2 Datasets

The provided training dataset consists of a total of 7899 answers, split into train and trial data and covering a total of 78 indivdidual questions. We deemed that appropriate to train question-specific models. All answers came from a real-world ITS system, thus label distributions where unsurprisingly uneven, even including a small number of answers with only correct or only incorrect answers in the dataset.

3 Method

Our system explores a number of complementary approaches: For the *unseen answers* track, we compare both instance-based and similarity-based approaches (Horbach and Zesch, 2019). For the *unseen questions* track, we experiment with approaches using generative LLMs.

In all experiments presented here, we use the train-trial data split provided by the Shared Task organizers. We evaluate using quadratically weighted kappa (QWK) where - although we in some in-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1201–1209

stances train one model per question - kappa values are computed across the whole trial data (instead of computing QWK prompt-wise and then averaging these individual values).

3.1 Similarity-Based Scoring

We tested two similarity-based scoring setups, Label-Aligned Similarity Learning (**LaSiLearn**) (Bexte et al., 2022) and **SetFit** (Tunstall et al., 2022). Both approaches are designed to work with SBERT (Reimers and Gurevych, 2019) base models. We ran both LaSiLearn and SetFit with three different multilingual base models, namely *paraphrase-multilingual-mpnet-base-v2* (**mpnet**), *paraphrase-multilingual-MiniLM-L12-v2* (**mini-lm**), and *distiluse-base-multilingual-cased-v1* (**distil**). All models were used via Hugging Face.

LaSiLearn LaSilearn adapts SBERT models so that the cosine similarity of answer embeddings reflects the similarity of the scores of two answers. For training supervision, pairs of two answers are built from the training data and assigned a similarity label that reflects whether the two answers in a pair share the same (similarity label 1) or a different (similarity label 0) score. With a fine-tuned model, test answers can then be compared to training answers. The score for which training answers have the highest average similarity to a test answer is then used as the prediction. For example, if the average similarity of a test answer is .4 for comparing it to *incorrect* training answers, .6 for *partially correct* training answers and .*8* for *correct* training answers, the model would predict the answer to be *correct*. LaSiLearn inference is also possible directly with the pretrained model, without any fine-tuning. Thus, we also test performance of the pretrained base models. As previous work had shown that for pretrained models it is best to predict the score of the most similar answer (Bexte et al., 2022, 2023), we adopt this strategy as well.

SetFit SetFit is a partially similarity-based scoring approach designed for smaller volumes of training data. First, the SBERT model is tuned as described for LaSiLearn in the previous section. Thereby, the embedding space becomes more representative of the answer space. In a second training loop, a classification head is then added to the model and fine-tuned using the same training data that was used for the similarity-based training. At inference, predictions are made using this classification head.

Task	mpnet	mini-lm	distil	vote	vote-sim
2-way					
$\text{LaSiLearn}_{\text{pre}}$	.50	.45	.50	.51	.49
LaSiLearn	.59	.61	.59	.63	.60
SetFit	.65	.62	.62	.65	–
3-way					
$\text{LaSiLearn}_{\text{pre}}$	.59	.54	.61	.61	.58
LaSiLearn	.71	.71	.71	.74	.75
SetFit	.70	.68	.71	.70	–

Table 1: QWK results for similarity-based scoring on trial data. $\text{LaSiLearn}_{\text{pre}}$ indicates direct use of the pretrained base model without any fine-tuning.

Hyperparameters For both LaSiLearn and SetFit, we set the same hyperparameters. We train with a batch size of 16 for at most 1500 steps, evaluating performance on validation data every 100 steps. The model that performs best on the validation data is kept for final evaluation on the test data. 10% of the training data for each prompt were set aside as validation data.

Results Table 1 shows performance of the similarity-based models on the trial data. Since we ran both scoring methods with three separate multilingual base models, we can aggregate their predictions via voting. For both LaSiLearn and SetFit, we performed majority voting (**vote** in Table 1). In the case of LaSiLearn, we also ran voting based on the similarities the model found. We took the prediction of the model that based its prediction on the highest similarity value (**vote-sim** in Table 1). For example, if one model predicted an answer to be *partially correct* based on a similarity of .78 to the reference answers and another said it would be *incorrect*, but based this on a similarity of .80, we adopt the prediction of the model that found the higher similarity and thus predict the answer to be *incorrect*.

In 2-way scoring, SetFit slightly outperforms LaSiLearn. The mpnet base model performs best, on par with a voted average of the three base models. For LaSiLearn, voting increases performance from .61 QWK to .63. On the 3-way data, LaSiLearn performs better than SetFit. Again, performance is pushed by voting, from .71 QWK to .75 with simlarity-based voting. For SetFit, the distil base model performs best, outperforming the voted average.

Based on the results, we chose the following configurations for our submission on the final test data:

- 2-way LaSiLearn: vote
- 3-way LaSiLearn: vote-sim
- 2-way SetFit: mpnet
- 3-way SetFit: distiluse

3.2 Instance-Based Scoring

In the instance-based scoring approach, we treat the task as a supervised classification problem, where models learn to score the correctness of answers from lexical material in the training data. The results are shown in Table 2.

3.2.1 Shallow Classifiers

As a baseline, we trained different shallow classifiers (Random Forest (RF), Support Vector Classifier (SVC), Logistic Regression (LR)) with unigram and bigram features. In Table 2, we only report LR, as this model performed best. We used the scikit-learn (Pedregosa et al., 2011) implementation and apart from setting *max_iter* to 1000 left all parameters at their respective default values.

3.2.2 Transformer Models

We fine-tuned three distinct pre-trained transformer-based models. All models were used via Hugging Face.

- BERT: The standard multilingual BERT base model. (Devlin et al., 2019)
- German-BERT: A BERT model specifically pre-trained on large-scale German corpora.[1]
- G-EdSciBERT: A specialized German model focused on educational and scientific domains. (Latif et al., 2024)

Experimental Setup All three models were fine-tuned using the same hyperparameter configuration to ensure a fair comparison. We utilized a learning rate of $2e-5$, a batch size of 7, and trained for 6 epochs. The models were trained on the provided training set without validation and evaluated on the trial data to determine their performance for the 2-way and 3-way classification tasks.

Results Table 2 summarizes the performance of these models on the trial data, reported using the Quadratically Weighted Kappa (QWK) metric. The trial results demonstrate that G-SciEdBERT consistently outperforms the other models on both tasks, achieving QWK values of 0.62 for 2-way classification and 0.67 for 3-way classification. This suggests that the domain-specific pre-training on German scientific and educational texts provides a significant advantage for the SAS task.

Task	LR	BERT	German BERT	G-SciEdBERT
2-way	.52	.47	.58	.62
3-way	.63	.60	.65	.67

Table 2: QWK results for instance-based models on trial data.

Prompt-Specific vs. Generic Training We further investigated the impact of prompt-specific training compared to a generic model approach. In the Generic Model (GM) setup, a single model is trained on the entire pooled dataset. In contrast, we trained individual Prompt-Specific Models (PSM) for each unique question to capture prompt-specific nuances. Our comparison, unsurprisingly, revealed a performance trade-off with the volume of available training data: PSMs have a comparative performance to GMs when the training set for a given prompt exceeds 100 instances. In this range, the PSM has a 52.2% probability of outperforming the GM. When the sample size is small, the GM tends to be more robust. The shared knowledge across prompts in the GM yields a higher average Kappa value compared to PSMs, which likely suffer from overfitting on limited instances. Based on the results, we chose generic training of G-SciEdBERT for our submission on the final test data.

Augmented Input We also experimented with augmenting the training input by concatenating the question text and the scoring rubric to the student answers. We hypothesized that this would provide the model with a better basis for determining correctness. However, this approach did not yield a consistent improvement, potentially due to increased noise or input length limitations.

3.3 LLM-based Scoring

Both closed- and open-source models were evaluated in a zero-shot setting, following the same general prompting strategy: each model received the question, the scoring rubric, and the student answer, and was instructed to predict a score label.

[1] https://huggingface.co/google-bert/bert-base-german-cased

Task	GPT5-mini	Phi-4	Gemma 3
2-way	.50	.46	.49
3-way	.39	.49	.51

Table 3: QWK results for LLMs on trial data.

3.3.1 Closed Source Models

We evaluated `gpt-5-mini` via the OpenAI API with default model parameters, using task-specific prompts for the 2-way and 3-way settings (see Figure 1 and 2 in the Appendix, respectively). Table 3 shows that `gpt-5-mini` achieves a QWK of .50 on the 2-way task and .39 on the 3-way task. Performance was measured on the trial data. The 2-way task is comparable to the weaker instance-based models, yet drops substantially for the 3-way task.

3.3.2 Open Source Models

For the open-source models, the same prompt was used for both 2-way and 3-way scoring, since the 2-way rubric subsumes the *partially correct* category under *incorrect*. For the full prompt, see Figure 3 in the Appendix. If a model failed to produce a valid response (i.e. one of the labels), it was re-prompted. If it still failed to answer after it had been prompted 10 times, the prediction was inserted as *invalid*. For evaluation, we took a permissive approach and treated answers with an *invalid* prediction as *correct*.

The two open models we prompted to score answers were Phi-4 (Abdin et al., 2024) and Gemma 3 (27b) (Team, 2025), as these were the two models that performed best in preliminary tests. Both models were accessed via the Ollama API on a local server with default parameters. Table 3 shows model performance on the trial data. Gemma 3 performs better than Phi-4 on both the 2-way and 3-way data. It is just .01 behind GPT5-mini for the 2-way data and outperforms it by .12 QWK on the 3-way data.

For submission on the final test data, we decided on the two open source models, since especially Gemma 3 performed comparably well in both the 2-way and the 3-way setting.

3.4 Influence of Normalization

As part of the similarity-based and instance-based approaches, we also conducted a set of experiments where we used normalized versions of the student answers rather than the raw versions. For normalization, we used TransGEC (Fang et al., 2023) and `MRNH/mbart-german-grammar-corrector`[2], two grammatical error correction systems trained on the Falko-MERLIN dataset for German (Boyd, 2018). In the similarity-based setting, the TransGEC normalization was able to outperform a model using pretrained SBERT on the raw version of the trial data both in the 2-way and 3-way settings. However, when SBERT was finetuned, then normalization proved to have no benefit. In the instance-based approach, the use of normalization mostly proved deleterious, sometimes notably so. Normalization was thus not used in any of the officially submitted runs.

3.5 Classifier Ensembles

Ensemble methods often improve robustness by combining complementary model predictions. This is particularly relevant in our setting, where the individual models differ substantially in how they arrive at their decisions: similarity-based models compare answers to previously seen responses, instance-based models learn direct label mappings, and LLMs rely on rubric-guided prompting.

We therefore evaluated majority-vote ensembles across different classifier combinations. To combine predictions in the 3-way setting, we map labels to numeric values (*incorrect* = 0.0, *partially correct* = 0.5, *correct* = 1.0), average the predictions across models, and round to the nearest valid label.

We evaluated all candidate classifier combinations on the trial data and selected the best-performing ensembles for the official submissions. Overall, the best ensembles consistently outperformed the strongest individual models, indicating that the approaches provide complementary information. Our best ensemble, consisting of *LaSiLearn (mpnet)*, *LaSiLearn$_{pre}$ (distil)*, *SetFit (distil)*, *G-SciEdBERT*, *Phi-4*, and *GPT5-mini* for the 3-way task and *LaSiLearn (mini-lm)*, *SetFit (mpnet)*, and *GSciEdBert* for the 2-way task yielded an improvement from 0.710 on the best individual model to 0.783 QWK on the trial data in the 3-way task and from 0.654 to 0.679 on the 2-way task .

4 Results

Tables 4 (unseen answers) and 5 (unseen questions) present results on the test data of the shared Task. We report Quadratically Weighted Kappa (QWK).

[2] https://huggingface.co/MRNH/mbart-german-grammar-corrector

Model	2-way	3-way
G-SciEdBERT	.59	.64
LaSiLearn	.62	.71
SetFit	.64	.71
best-1	.65	.74
best-2	.66	.76
best-3	.64	.74

Table 4: Evaluation Results for Unseen Answers. Note that best-1 to best-3 are the best-performing classifier combinations on the trial data and are different between 2-way and 3-way.

Model	2-way	3-way
G-SciEdBERT	.36	.50
Phi-4	.38	.51
Gemma 3	.39	.52
no_gpt	.42	.58
no_instance	.45	.52
no_gemma	.45	.54
no_phi	.46	.54

Table 5: Evaluation Results for Unseen Questions

Unseen Answers Much like on the trial data, instance-based training of G-SciEdBERT is outperformed by the similarity-based models. Between the two similarity-based models, SetFit is again slightly ahead of LaSiLearn on the 2-way data, while both models perform on par on the 3-way data. Our ensemble strategy is able to push performance for both the 2-way and 3-way setting, to .66 and .76, respectively. We submitted the five best-performing runs on the trial data and report the results for the first three of them in Table 4. For details on which models are part of these ensembles, see Table 6 in the appendix. On the test data, the second best run of these five yielded the best performance on the evaluation data, hinting at a certain overfitting in our model selection process. The best-performing submission (best-2) ranked 5th (2-way) and 4th (3-way) in the shared task.

Unseen Questions In addition to the LLMs, we also used the instance-based G-SciEdBERT for prediction on the unseen questions. It falls slightly behind LLM performance, but only by at most .03 QWK on the 2-way data and .02 QWK on the 3-way data. In both settings, Gemma 3 performs best, achieving QWK .39 on 2-way data and QWK .52 on the 3-way data. We also tested voting combinations with three out of the four models (G-SciEdBERT and the three LLMs). These results are included as *no-** in Table 5, indicating which of the four models was *not* part of the voting. In general, voting again increases performance. Curiously, dropping GPT5-mini leads to the lowest increase in performance on 2-way data, but the best result of QWK .58 on the 3-way data. This is likely because gpt already performed poorly on the 3way trial data, but we dont have the numbers for test data. The best result on the 2-way data (QWK .46) is obtained when Phi-4 is not included in the voting. With respect to performance relative to the other submissions to the shared task, we placed 6th in the 2-way condition (no_phi) and 4th in the 3-way condition (no_gpt).

Notably, all models performed significantly better on the 3-way task than the 2-way task. This may be explained by the label distribution: in the 3-way task, the classes are relatively balanced (29.2% Correct, 36.3% Partially Correct, and 34.5% Incorrect). However, in the 2-way task, merging the 'partially correct' and 'incorrect' categories creates a significant class imbalance, with the 'incorrect' category accounting for 70.8% of the data. Separating "partially correct" as its own label in the 3-way task appears to provide a clearer signal for the models to learn.

5 Error Analysis: Influence of 'incoherent' scoring rubrics

A recurring source of confusion during manual inspection of the data was the missing task context: The shared task data originates from a realistic ITS context, but were provided without full context information about the surrounding material. Therefore, a number of questions were not fully transparent such as a question simply stating "Erkläre, in eigenen Worten, warum du dich für diese Lampe entschieden hast." (*Explain in your own words why you have chosen that specific lamp!*), with a rubric for a correct answer stating "Die SuS formulieren eine Begründung basierend auf physikalischen Prinzipien und gesellschaftlichen/ökologischen Beweggründen." (*The students formulate a justification based on physical principles and societal / ecological reasons.*). Without any context it can be hard for humans to determine whether an answer to that question is actually correct.

In an error analysis, we manually annotated such incomplete questions (16 in total) and then compared model performance on these incomplete questions to performance on the remaining trial set. However, QWK values did not substantially differ, hinting at the models capability to infer relevant information from either sample answers or rubric contexts.

6 Conclusion

In this paper, we compared various different approaches to short-answer scoring on a German science question dataset. We found that established instance- and similarity-based methods outperform approaches based on generative LLMs that take only the scoring rubric into account in the unseen answers track, whereas LLM-based approaches where the only applicable option in the unseen questions track. In both cases, we benefit from using an ensemble of classifiers instead of a single classifier only.

Acknowledgements

This work was conducted as part of the DFG-funded ASLAN project (Fördernummer: 563947383).

Limitations

This shared task treated classifier performance as the sole evaluation metric. In a practical application of Short Answer Scoring, however, considerations of robustness and fairness (Loukina et al., 2019; Schaller et al., 2024) can be equally important, particularly with respect to potential biases affecting specific subgroups of learners and should thus also be considered in evaluation.

Ethical Considerations

This work relies on a dataset provided by the shared task organizers. As such, data collection, anonymization, and privacy protection measures were handled prior to our use of the data.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

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.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring-how to make s-bert keep up with bert. In *Proceedings of the 17th workshop on innovative use of NLP for building educational applications (BEA 2022)*, pages 118–123.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2023. Similarity-based content scoring-a more classroom-suitable alternative to instance-based scoring? In *Findings of the association for computational linguistics: Acl 2023*, pages 1892–1903.

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

Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Fang, Xuebo Liu, Derek F. Wong, Runzhe Zhan, Liang Ding, Lidia S. Chao, Dacheng Tao, and Min Zhang. 2023. TransGEC: Improving grammatical error correction with translationese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3614–3633, Toronto, Canada. Association for Computational Linguistics.

Sebastian Gombert, Zhifan Sun, Fabian Zehner, Jannik Lossjew, Tobias Wyrwich, Berrit Katharina Czinczel, David Bednorz, Sascha Bernholt, Knut Neumann, Ute Harms, Aiso Heinze, and Hendrik Drachsler. 2026. Report on the bea 2026 shared task on rubric-based short answer scoring for german. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.

Andrea Horbach and Torsten Zesch. 2019. The Influence of Variance in Learner Answers on Automatic Content Scoring. *Frontiers in Education*, 4.

Ehsan Latif, Gyeong-Geon Lee, Knut Neumann, Tamara Kastorff, and Xiaoming Zhai. 2024. G-sciedbert: A contextualized llm for science assessment tasks in german. *arXiv preprint arXiv:2402.06584*.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 1–10.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on german learner essays from secondary education. In *Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea 2024)*, pages 210–221.

Gemma Team. 2025. Gemma 3 technical report.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

A Appendix

As supplementary material, we include some information that is intended to increase the transparency of our results.

LLM Prompts In Figures 1 (GPT-5 mini, 2-way scoring), 2 (GPT5-mini, 3-way scoring), and 3 (open source LLMs), we include the prompts that were used to score answers with LLMs.

Ensemble Voting In Table 6, we detail which models were part of our best voted scoring runs on the trial data.

```
Sie sind ein Experte fuer paedagogische Diagnostik im Bereich der deutschen
MINT-Bildung. Ihre Aufgabe ist es, Schuelerantworten basierend auf einer
vorgegebenen Rubrik zu bewerten.

Sie erhalten:
1. Eine Frage (auf Deutsch)
2. Eine Schuelerantwort (auf Deutsch)
3. Eine Bewertungsrubrik mit Kriterien fuer "Richtig" (Correct) und
   "Falsch" (Incorrect)

Ihre Aufgabe:
- Lesen Sie die Frage sorgfaeltig durch und verstehen Sie die Aufgabenstellung.
- Analysieren Sie die Antwort des Schuelers.
- Wenden Sie die Rubrik-Kriterien strikt an, um zu entscheiden, ob die
  Antwort "Correct" oder "Incorrect" ist.
- Beruecksichtigung von Teilverstaendnis: Wenn die Rubrik "Correct" als
  umfassendes oder vollstaendiges Verstaendnis definiert, muss eine Antwort,
  die nur Teilaspekte abdeckt, als "Incorrect" bewertet werden.

WICHTIG:
- Antworten Sie NUR mit dem Label: entweder "Correct" oder "Incorrect".
- Geben Sie keine Erklaerungen, Begruendungen oder zusaetzlichen Text aus.
- Die Bewertung muss sich strikt nach den Kriterien der Rubrik richten.
```

Figure 1: Prompt for 2-way classification with GPT5-mini.

Configuration	2-way Mix	3-way Mix
best-1	*LaSiLearn (mini-lm)*, *SetFit (mpnet)*, *GSciEdBert*	*LaSiLearn (mpnet)* *LaSiLearn$_{pre}$ (distil)*, SetFit (distil), *G-SciEdBERT*, *Phi-4*, *GPT5-mini*
best-2	*LaSiLearn (mini-lm)*, *LaSiLearn$_{pre}$ (distil)*, *SetFit (mpnet)*, *Gemma 3*, *GPT5-mini*	*LaSiLearn (mini-lm)*, *LaSiLearn (mpnet)*, *LaSiLearn$_{pre}$ (distil)*, *SetFit (distil)*, *G-SciEdBert*, *RF*, *Phi-4*, *GPT5-mini*
best-3	*LaSiLearn$_{pre}$ (mpnet)*, *SetFit (mpnet)*, *G-SciEdBERT*	*LaSiLearn (distil)*, *LaSiLearn (mpnet)*, *LaSiLearn$_{pre}$ (distil)*, *G-SciEdBERT*, *Phi-4*, *GPT5-mini*

Table 6: Details on the best configurations for voting on the trial data.

```
Sie sind ein Experte fuer paedagogische Diagnostik im Bereich der deutschen
MINT-Bildung. Ihre Aufgabe ist es, Schuelerantworten basierend auf einer
vorgegebenen Rubrik zu bewerten.

Sie erhalten:
1. Eine Frage (auf Deutsch)
2. Eine Schuelerantwort (auf Deutsch)
3. Eine Bewertungsrubrik mit Kriterien fuer "Richtig" (Correct),
   "Teilweise richtig" (Partially Correct) und "Falsch" (Incorrect)

Ihre Aufgabe:
- Lesen Sie die Frage sorgfaeltig durch und verstehen Sie die Aufgabenstellung.
- Analysieren Sie die Antwort des Schuelers.
- Wenden Sie die Rubrik-Kriterien strikt an, um zu entscheiden, ob die Antwort
  "Correct", "Partially Correct" oder "Incorrect" ist.
- Unterscheidung der Stufen: Vergleichen Sie die Antwort sorgfaeltig mit allen
  drei Rubrik-Stufen. Eine Antwort ist nur "Correct", wenn sie die Kriterien
  fuer vollstaendige Richtigkeit erfuellt. Deckt sie nur Teilaspekte ab, ist
  sie als "Partially Correct" zu bewerten. Fehlt jeglicher inhaltlicher Bezug
  zu den Rubrik-Kriterien, ist sie "Incorrect".

WICHTIG:
- Antworten Sie NUR mit einem der folgenden Labels: "Correct",
  "Partially Correct" oder "Incorrect".
- Geben Sie keine Erklaerungen, Begruendungen oder zusaetzlichen Text aus.
- Die Bewertung muss sich strikt nach den Kriterien der Rubrik richten.
```

Figure 2: Prompt for 3-way classification with GPT5-mini.

```
You are a teacher grading student answers for the following question:
Question: {question}
Answers are evaluated with the following rubric:
- {rubric.correct} (correct)
- {rubric.partial} (partially correct)
- {rubric.incorrect} (incorrect)
Please state which label you are assigning the following answer, based on the rubric:
Answer: {answer}
Answer only with the label you are assigning to the answer, without any additional explanation.
```

Figure 3: Prompting strategy used for evaluation of answers with open-source LLMs.