SDPA at BEA 2026 Shared Task 2: Efficient LLM Fine-Tuning for Rubric-based Short Answer Scoring

Zhexiong Liu[1*], **Jing Zhang**[2*]
[1]University of Pittsburgh, Pittsburgh, Pennsylvania 15260 USA
[2]Emory University, Atlanta, Georgia 30322 USA
zhexiong@cs.pitt.edu, jingz1@alumni.emory.edu

Abstract

Automated short-answer scoring (ASA) is an important yet challenging task in educational assessment as it aims to evaluate open-ended student responses against predefined scoring rubrics that are often interrelated. Although large language models (LLMs) have demonstrated impressive capabilities in text understanding and reasoning, their application to ASA has primarily focused on prompt-based inference, largely due to the limited availability of annotated data required for effective model training. In this work, we investigate parameter-efficient fine-tuning strategies for LLMs using ASA annotations in German. Our experiments show that fine-tuned LLMs consistently outperform both prompt-based and ensemble-based language models, suggesting domain-adaptive LLM fine-tuning is more effective than prompting alone for ASA.

Question: Assess the rate of spoilage when the apples are cut (Note: if you would like to include technical information on the chemical processes of spoilage, you can research and include it, citing your sources).

I. **Correct Answer**: The smaller the apples are cut, the larger their surface area that can react. The larger the surface area, the more collisions and therefore more effective collisions can occur, and the faster the reaction proceeds.

II. **Partially Correct Answer**: The larger the surface area of the apple, the more substances can react on the surface, e.g., oxidize (i.e., turn brown).

III. **Incorrect Answer**: Apples spoil faster at higher temperatures because the reaction rate is increased.

I. **Correct Rubric**: The students comprehensively assess the perishability of the apples based on the collision model.

II. **Partially Correct Rubric**: The students partially assess the perishability of the apples based on the collision model.

III. **Incorrect Rubric**: The students do not assess the perishability of the apples based on the collision model.

Figure 1: A short-answer scoring example where a single question could receive correct, partially correct, and incorrect answers based on the given rubrics. The original example is in German (Gombert et al., 2026).

1 Introduction

Student responses to open-ended domain-specific questions are often regarded as important indicators of their acquaintance with teacher-instructed content (Berliner and Rosenshine, 2017). Faithful assessment of their responses is essential for determining how well students have internalized curricular materials. In practice, however, manually evaluating short-answer responses remains dominant in classrooms, and it is both time-consuming and cognitively intensive since the responses are often diverse and a single question may accept multiple answers from different conceptual perspectives, which makes the manual scoring exceedingly difficult to scale. For example, Figure 1 shows three student answers scored as correct, partially correct, and incorrect based on highly interrelated rubrics. These challenges have thus motivated growing interest in the development of automated short-answer scoring (ASA) systems that can provide rubric-aligned assessments with minimal human burdens. Early automated methods focused on either lexical and semantic text matches (Ramachandran et al., 2015), or training Transformer-based S-BERT (Bexte et al., 2022). Although these models achieve reasonable performance, they have limited capability to align complex student responses with nuanced rubrics, particularly when training data are scarce. Recent large language models (LLMs) have been used to solve several challenging problems in the educational domain, e.g., essay revision (Li et al., 2024), propositional logic (Liu et al., 2024b), and mathematical geometry (Zhang et al., 2024). However, existing LLM approaches to ASA have predominantly relied on prompt methods (Henkel et al., 2024; Ferreira Mello et al., 2025), which tend to underperform when several rubrics are interrelated, e.g., partially correct versus correct rubrics. The potential of fine-tuned LLMs for ASA, particularly under low-resource conditions, therefore

[*]These authors contributed equally to this work

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1244–1251

warrants further exploration. In this study in particular, we investigate parameter-efficient fine-tuning (PEFT) strategies to align LLMs with the linguistic features and rubric-matching requirements of short-answer scoring tasks.

We present our system submissions to the ASA shared task co-located at the Building Educational Applications (BEA) workshop (Gombert et al., 2026). The shared task aims to contribute to non-English benchmark datasets with four submission tracks described by two dimensions: whether the answers or questions are unseen at test time, and whether scoring follows a three-way (correct, partially correct, incorrect) or two-way (correct, incorrect) scheme. The two-way tracks collapse partially correct responses into the incorrect class, thereby posing an additional challenge for interpreting interrelated rubrics. We first establish a strong baseline using small pre-trained language models (Liu et al., 2019; Devlin et al., 2019; Chan et al., 2020), and then investigate an ensemble strategy that integrates small language models with LLMs to exploit their complementary strengths. In addition, we investigate PEFT (Hu et al., 2022; Liu et al., 2024a) strategies for LLM fine-tuning to assess their task-specific adaptation under low-resource conditions. Our system achieves second place on the binary unseen-question track and third place on the remaining three tracks. These findings demonstrate that LLMs can be effectively fine-tuned for ASA even with limited labels, providing a practical solution for low-resource educational assessment.

2 Related Work

Small Language Models. In the literature, ASA mostly relied on hand-crafted features for similarity measurement between answers and rubrics, e.g., semantic-based dependency graph alignment (Mohler et al., 2011), machine learning-based methods (Burrows et al., 2015), and several scoring systems described in the SemEval-2013 Task 7 shared task (Dzikovska et al., 2013). In addition, pre-trained Transformer encoders have substantially advanced ASA tasks (Haller et al., 2022), e.g., domain-adaptive BERT (Devlin et al., 2019) trained with augmenting data from the domain-specific textbooks (Sung et al., 2019), SBERT (Reimers and Gurevych, 2019) based on transfer learning and data augmentation (Bonthu et al., 2023). Despite these advances, small encoder models have limited capacity to handle nuanced open-ended responses and often struggle to distinguish between incorrect and partially correct answers. This limitation has motivated increasing interest in leveraging LLMs, which offer stronger semantic understanding and reasoning capabilities.

Large Language Models. Recent work has applied large language models (LLMs) to ASA scoring. Henkel et al. (2024) suggest that LLMs with few-shot prompts achieve human scoring in K-12 education, while Chamieh et al. (2024) and Ferreira Mello et al. (2025) indicate that prompt-based LLMs are, in contrast, less competitive than fine-tuned BERT approaches, but LLMs show promising capabilities when configured with optimized prompts such as few-shot examples and clear instructions. These observations are also presented while assessing student explanations, where fine-tuned FLAN-T5 (Chung et al., 2024) outperforms LLMs in-context learning (Carpenter et al., 2024) as well as reasoning propositional logic, where a BERT-based model outperforms Llama-based LLMs (Liu et al., 2024b). Thus, these findings reveal the limited domain generalizability of LLM prompting methods and have thus motivated the development of more task-adaptive LLM solutions. More recent ASA work investigates a QLoRA-based cross-prompt fine-tuning method (Funayama et al., 2025) and a LLM fine-tuned framework with generated feedback (Aggarwal et al., 2025). However, rubric-aware LLM fine-tuning with limited data, particularly in non-English settings (Padó et al., 2024), remains underexplored, which motivates our study.

3 Dataset

The ASA scoring task is based on a novel dataset, ALICE-LP-1.0 (Gombert et al., 2026), which contains a German-language corpus of mid- and high-school level answers to questions across four STEM domains: physics, mathematics, biology, and chemistry. Students answered short-answer items embedded in synchronous Moodle-based learning activities under teacher supervision. Each instance consists of a question, a student response, a textual rubric defining the scoring rubrics, and a label on a three-way scale (*incorrect*, *partially correct*, *correct*), as well as a binary version that collapses *partially correct* into *incorrect* label (see Figure 1). The training set contains 7,899 responses, the trial set used for validating the models contains 827 responses, and the evaluation set used for

leaderboard ranking is split into two subsets: 2,008 *unseen answers* responses obtained during training and 3,168 *unseen questions* responses held out from training. The evaluation design follows the protocol established by SemEval-2013 (Dzikovska et al., 2013) and SAF (Filighera et al., 2022). An example of the ASA data is shown in Figure 1.

4 Method

We address the ASA task with a batch of small and large language models from two perspectives. **Ensemble LLMs**, as described in Section 4.1, first employ small language model encoders to learn task-specific representations from labeled responses and produce calibrated predictive probability distributions of the candidate labels. Afterward, we ensemble multiple fine-tuned models, including few-shot predictions from LLMs, i.e., Claude-Opus-4.6 (Anthropic, 2026), to provide complementary scoring predictions. The final results are based on ensemble probabilities across candidate models. In addition, **Finetuned LLMs**, as described in Section 4.2, leverage LLM layers and a classifier head for fine-tuning using parameter-efficient methods, e.g., LoRA (Hu et al., 2022) and DoRA (Liu et al., 2024a). Unlike most generative LLMs that aim to generate predictive labels by learning the next-token probabilities, our fine-tuned LLMs instead learn to directly predict probabilities over the scoring labels. The two approaches are complementary as the encoder ensembles excel when training questions overlap with test questions (unseen answers), while the instruction-tuned LLM generalizes better to novel questions (unseen questions) by leveraging its strong linguistic capabilities.

4.1 Encoder-Ensemble LLMs

We preprocess the input text by concatenating the question, student answer, and rubric into a single sequence, i.e., *Question: {question} [SEP] Answer: {answer} [SEP] Rubric: {label$_1$}: {desc$_1$} | {label$_2$}: {desc$_2$} | {label$_3$}: {desc$_3$}*. This format provides the model with the full context needed to assess the answer against the rubric criteria, inspired by prior ASA scoring approaches (Sung et al., 2019). We train three Transformer-based models, ranging from monolingual German to multilingual language models:

- **GELECTRA-large** (Chan et al., 2020): A German ELECTRA model pre-trained on a large German corpus using the replaced token detection objective. It has demonstrated strong performance on GermEval14 (Benikova et al., 2014) and GermEval18 (Wiegand et al., 2018).
- **XLM-RoBERTa-large** (Conneau et al., 2020): A multilingual model pre-trained on 100 languages using masked language modeling on 2.5TB of clean CommonCrawl data.
- **BERT-base-german-cased** (Chan et al., 2020): A German BERT model pre-trained on German Wikipedia and news corpora.

We combine predictions from the trained encoder models with few-shot predictions from Claude-Opus-4.6 (Anthropic, 2026), without fine-tuning. Claude-Opus-4.6 uses a small number of training examples as demonstrations, of which the same questions are excluded for unseen answer tracks. It returns a score label based on its rubric interpretation and reasoning capabilities. The final ensemble uses soft voting with optimized weights. Each fine-tuned encoder produces a softmax probability distribution over labels, while Claude-Opus-4.6's predictions are regarded as hard labels (one-hot). The ensemble prediction is:

$$\hat{y} = \arg\max_{c} \sum_{i=1}^{n} w_i \cdot p_i(c \mid x) \qquad (1)$$

where w_i are model weights optimized via the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) to maximize QWK on the trial set, p_i is the predicted probability for label c, and n is the number of ensembled models. Note that models that fail to learn (QWK $= 0$ on the trial set) are excluded from the ensemble for that track. We also explore multiple fusion ratios between the encoder ensemble and Claude-Opus-4.6 predictions (e.g., 70% encoder and 30% Claude-Opus-4.6), as well as a confidence-based selection strategy in which the encoder ensemble is applied when its maximum softmax probability exceeds a confidence threshold (e.g., 0.7), and Claude-Opus-4.6 is used for the remaining uncertain cases.

Implementation Each model is trained with random seeds to improve robustness through diversity, yielding up to 6 model variants per track. Also, all models are trained using the HuggingFace[1] with a linear classification head. The Calude-Opus-4.6 uses 0 temperature, and a maximum token of 50. BERT-based model training hyperparameters

[1] https://huggingface.co/

Hyperparameter	Value
Batch size	16
Learning rate (encoder)	2×10^{-5}
Optimizer	AdamW
Max epochs	8
Max sequence length	512
LR scheduler	warmup (200 steps)
Loss function	CE with class weights

Table 1: Training hyperparameters for encoder models.

are summarized in Table 1. We also employ differential learning rates (Cazé and van der Meer, 2013), applying a $5\times$ multiplier to the classification head compared to the pre-trained encoder layers. The inverse-frequency class weights in the cross-entropy loss are used to address label imbalance, which is particularly important for the 2-way track where the Incorrect class comprises 71% of samples in the training set.

4.2 Instruction-Finetuned LLMs

While encoder models are effective for classification, they take ASA scoring as pattern matching rather than explicitly rubric-based reasoning. While full fine-tuning of LLMs can address this limitation, it is prohibitively expensive and requires large-scale annotated data. We thus leverage PEFT methods (Hu et al., 2023) to adapt low-resource fine-tuning while preserving their strong linguistic capabilities. Unlike ensemble-LLMs (Sec. 4.1), which are built from multiple models, fine-tuned LLMs produce a standalone scorer directly for inference. Specifically, we use Llama-3.1-8B (Grattafiori et al., 2024) as our base model, as it has strong multilingual capabilities and instruction-following performance. The model is instruction-tuned on the scoring task, using the instruction formatted following the prior Alpaca instruction template[2] (Taori et al., 2023).

```
### Instruction:
Score the following student response based on
the given rubric.
### Input:
Question: {question}
Student Response: {answer}
Rubric:
- Correct: {rubric_correct}
- Partially correct: {rubric_partial}
- Incorrect: {rubric_incorrect}
```

Inspired by previous PEFT work on text classification (Liu, 2025; Liu and Litman, 2025), we employ adapter-based methods for efficient LLM fine-tuning, i.e., LoRA (Hu et al., 2022), which freezes the pre-trained model weights and injects trainable low-rank decomposition matrices into each transformer layer. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the modified forward pass is:

$$h = W_0 x + \Delta W x = W_0 x + BAx \qquad (2)$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d, k)$. This reduces trainable parameters by orders of magnitude while preserving the model's pre-trained knowledge. In addition, we also used the recent DoRA (Liu et al., 2024a), which extends LoRA by decomposing the pre-trained weight into magnitude and direction components:

$$W' = m \cdot \frac{W_0 + BA}{\|W_0 + BA\|_c} \qquad (3)$$

where $m \in \mathbb{R}^{1\times k}$ is a learnable magnitude vector and $\|\cdot\|_c$ denotes the column-wise norm. This decomposition enables independent adaptation of weight magnitude and direction, closely approximating full fine-tuning dynamics while maintaining the parameter efficiency of LoRA. DoRA has been shown to outperform LoRA on several NLP tasks (Liu et al., 2024a).

Hyperparameter	Value
Quantization	8-bit
LoRA/DoRA rank r	32
Alpha α	64
Target modules	q, k, v, o projections
Dropout	0.05
Learning rate	1×10^{-4}
Batch size	16
Max epochs	5
Max sequence length	512
LR scheduler	warmup (100 steps)

Table 2: Hyperparameters for LLMs with LoRA/DoRA.

4.2.1 Implementation

Specifically, we apply 8-bit quantization, following prior work (Liu and Litman, 2026), to reduce memory consumption. We build the framework using PyTorch[3] and HuggingFace, and optimize task losses using the Adam optimizer on an Nvidia A100. We fine-tune LLMs on the training set and tune hyperparameters on the trail set. At inference time, the fine-tuned Llama3.1-8B model outputs label probabilities given the instruction-formatted input. A softmax is applied to return the predicted label. The detailed hyperparameter settings are shown in Table 2.

[2]The instructions used in the experiment were in German, but are translated into English for visualization.

[3]https://pytorch.org/

Method	Model	QWK		W-Precision		W-Recall		W-F1	
		UA	UQ	UA	UQ	UA	UQ	UA	UQ
Ensemble-LLMs	BERT-base-de	0.5514	0.4035	0.8136	0.7519	0.8147	0.7592	0.8141	0.7545
	GELECTRA-large	0.6298	0.4154	0.8462	0.7583	0.8456	0.7673	0.8459	0.7606
	Optimized Ensemble	0.6516	0.4362	0.8556	0.7667	0.8575	0.7747	0.8563	0.7688
	Ensemble + Claude-Opus-4.6	**0.6617**	**0.4609**	**0.8598**	**0.7756**	**0.8615**	**0.7812**	**0.8604**	**0.7776**
Finetuned-LLMs	Llama3.1-8B + LoRA	**0.6822**	**0.5348**	**0.8685**	**0.8056**	0.8655	0.8036	**0.8667**	0.8045
	Llama3.1-8B + DoRA	0.6786	0.5307	0.8665	0.8056	**0.8660**	**0.8108**	0.8663	**0.8069**

Table 3: The 2-way track test results for Unseen Answers (UA) and Unseen Questions (UQ).

Method	Model	QWK		W-Precision		W-Recall		W-F1	
		UA	UQ	UA	UQ	UA	UQ	UA	UQ
Ensemble-LLMs	BERT-base-de	0.6667	0.4603	0.6893	0.5670	0.6688	0.5401	0.6727	0.5452
	XLM-R-large	0.6656	0.4652	0.6957	0.5819	0.6802	0.5476	0.6833	0.5521
	GELECTRA-large	0.7352	0.5145	0.7451	0.6015	0.7305	0.5755	0.7337	0.5774
	Optimized Ensemble	0.7414	0.5228	0.7488	0.6119	0.7345	0.5842	0.7377	0.5866
	Ensemble + Claude-Opus-4.6	**0.7566**	**0.5446**	**0.7695**	**0.6165**	**0.7524**	**0.5991**	**0.7554**	**0.6030**
Finetuned LLMs	Llama3.1-8B + LoRA	**0.7760**	**0.6444**	**0.7748**	0.6338	**0.7634**	**0.6325**	**0.7661**	**0.6330**
	Llama3.1-8B + DoRA	0.7556	0.6292	0.7454	**0.6440**	0.7460	0.6251	0.7452	0.6300

Table 4: The 3-way track test results for Unseen Answers (UA) and Unseen Questions (UQ).

4.3 Results

Regarding ensemble-LLMs for the two-way and three-way track performance in Table 3 and Table 4, GELECTRA-large consistently achieves the strongest individual performance, benefiting from German-specific pre-training. In terms of XLM-RoBERTa-large, it exhibits training fluctuation, with some seeds failing to converge (e.g., $\text{QWK} = 0$), particularly on the 2-way track, thus we did not submit its performance. In contrast, GELECTRA-large outperforms the larger multilingual XLM-RoBERTa-large in three-way track. This demonstrates that monolingual pre-training yields more useful representations than model size for language-specific classification tasks, whereas XLM-RoBERTa-large exhibits high sensitivity to experimental configuration, as some runs fail to learn in our pilot study. This might be caused by a linguistic adaption from its multilingual pre-training to the German classification objective, given the relatively small training set.

Regarding LLMs for the two-way and three-way scoring, ensemble Claude-Opus-4.6's few-shot predictions without fine-tuning on the task consistently provide a complementary strength by leveraging rubric-level reasoning. In other words, incorporating Claude-Opus-4.6's predictions into the encoder ensemble is particularly beneficial for ambiguous cases where discriminative models produce low-confidence outputs. In contrast, the instruction fine-tuned Llama3.1-8B models substantially outperform the ensemble-based method on both the unseen answer and unseen question tracks. Specifically, fine-tuning yields relatively larger gains on the 2-way track than on the 3-way track, suggesting that 3-way ASA scoring, which requires distinguishing partially correct from incorrect responses, is inherently more challenging. Although LoRA generally outperforms DoRA in terms of QWK, their weighted F1 scores remain comparable. Given that LoRA is also more computationally efficient, it is regarded as the preferred PEFT method. Notably, the fine-tuning approaches offer the advantage of a single scorer that evaluates responses against rubrics without relying on an ensemble of multiple models. This is particularly beneficial for the unseen questions track, where generalization to novel questions and rubrics is critical.

5 Conclusion

We presented a set of novel ensemble-based and fine-tuned LLM methods for ASA scoring in German. The encoder ensemble augmented with few-shot Claude-Opus-4.6 predictions demonstrates advantages over small language models, while the instruction-tuned Llama3.1-8B scorer exhibits superior performance in learning complex scoring patterns. Our results suggest that parameter-efficient fine-tuning enables LLMs to score diverse student responses even with limited training data. These findings offer a practical solution for deploying LLM-based ASA scoring systems in low-resource educational assessment.

Limitations

Our encoder ensemble optimization relies on the trial set for weight tuning, which may lead to overfitting on the evaluation set. In addition, the ensemble approach increases system complexity, requiring training multiple models and careful ensemble-based strategies. While fine-tuned LLMs with PEFT achieve substantial improvement, they require GPU resources, potentially limiting deployment in low-resource educational contexts. Furthermore, our experiments are conducted exclusively on German short-answer data, and the effectiveness of the proposed methods on other languages or domains remains unverified.

Ethics Statement

This work uses the ALICE-LP-1.0 dataset, which was collected and annotated for academic research purposes with appropriate institutional oversight. As the dataset contains student responses from minors in German secondary schools, we do not attempt to re-identify individual students. Our automated scoring systems are intended to assist, not replace, human graders; deployment in high-stakes assessment settings would require thorough validation and human oversight to ensure fairness across student populations. We acknowledge that both encoder models and LLMs may introduce biases during pretraining that could affect scoring equity. Additionally, LLM-generated scores may exhibit inconsistencies or errors that require careful human review prior to educational deployment. Lastly, the use of LLMs may raises practical concerns regarding cost, accessibility, and reproducibility.

References

Dishank Aggarwal, Pritam Sil, Bhaskaran Raman, and Pushpak Bhattacharyya. 2025. "i understand why i got this grade": Automatic short answer grading (asag) with feedback. pages 304–318.

Anthropic. 2026. Introducing claude opus 4.6.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014. Germeval 2014 named entity recognition shared task: companion paper. In *Workshop Proceedings of the 12th edition of the KONVENS conference*, pages 104–112.

David C Berliner and Barak Rosenshine. 2017. The acquisition of knowledge in the classroom. In *Schooling and the acquisition of knowledge*, pages 375–396. Routledge.

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.

Sridevi Bonthu, S. Rama Sree, and M.H.M. Krishna Prasad. 2023. Improving the performance of automatic short answer grading using transfer learning and augmentation. *Engineering Applications of Artificial Intelligence*, 123:106292.

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.

Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 403–413.

Romain D Cazé and Matthijs AA van der Meer. 2013. Adaptive properties of differential learning rates for positive and negative outcomes. *Biological cybernetics*, 107(6):711–719.

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)*, pages 309–315.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Alexis Conneau, Kartik Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Myroslava O 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.

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*, LAK '25, page 93–103, New York, NY, USA. Association for Computing Machinery.

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.

Hiroaki Funayama, Yuichiroh Matsubayashi, Yuya Asazuma, Tomoya Mizumoto, and Kentaro Inui. 2025. Cross-prompt pre-finetuning of language models for short answer scoring. *International Journal of Artificial Intelligence in Education*, pages 1–22.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 300–304, New York, NY, USA. Association for Computing Machinery.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanen Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. *Proceedings of ICLR*.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 5254–5276.

Tianwen Li, Zhexiong Liu, Lindsay Matsumura, Elaine Wang, Diane Litman, and Richard Correnti. 2024. Using large language models to assess young students' writing revisions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 365–380, Mexico City, Mexico. Association for Computational Linguistics.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024a. DoRA: Weight-decomposed low-rank adaptation. *Proceedings of ICML*.

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

Zhexiong Liu. 2025. *Understanding and Generating Text Revisions With Intention-Adaptive Large Language Models*. Ph.D. thesis, University of Pittsburgh.

Zhexiong Liu and Diane Litman. 2025. Efficient layer-wise LLM fine-tuning for revision intention prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15319–15334, Suzhou, China. Association for Computational Linguistics.

Zhexiong Liu and Diane Litman. 2026. Intention-adaptive LLM fine-tuning for text revision generation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1263–1281, Rabat, Morocco. Association for Computational Linguistics.

Zhexiong Liu, Jing Zhang, Jiaying Lu, Wenjing Ma, and Joyce C. Ho. 2024b. Logicprpbank: A corpus for logical implication and equivalence. In *Proceedings of the 2024 AAAI Conference on Artificial Intelligence*, volume 257 of *Proceedings of Machine Learning Research*, pages 57–65. PMLR.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762, Portland, Oregon, USA. Association for Computational Linguistics.

John A Nelder and Roger Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Ulrike Padó, Yunus Eryilmaz, and Larissa Kirschner. 2024. Short-answer grading for german: Addressing the challenges. *International Journal of Artificial Intelligence in Education*, 34(4):1321–1352.

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.

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.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Jiaxin Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Cheng-Lin Liu, and Yashar Moshfeghi. 2024. Geoeval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1258–1276.