WSE Research at BEA 2026 Shared Task 2: Multi-Strategy Rubric-Based Short Answer Scoring for German*

Jonas Gwozdz and Andreas Both
Leipzig University of Applied Sciences, Leipzig (Germany)
Web & Software Engineering Research Group (WSE Research)
{jonas.gwozdz, andreas.both}@htwk-leipzig.de

Abstract

We describe the WSE Research system for the BEA 2026 Shared Task 2 on Rubric-based Short Answer Scoring for German. Our system combines rubric-conditioned prompting with TF-IDF exemplar retrieval, LoRA fine-tuning of open-source Qwen models, and prediction aggregation across complementary scorers. The central question is when prompt engineering, parameter-efficient adaptation, and aggregation each help for rubric-based grading. On the ALICE-LP-1.0 trial set, a fine-tuned Qwen2.5-32B reaches QWK 0.769, surpassing the strongest prompted commercial baseline (Gemini 3 Flash, 0.748). On the official test set, the system ranks second on three tracks and third on the remaining one. Overall, the results show that rubric-conditioned fine-tuning is a competitive and cost-effective alternative to commercial APIs for German short answer scoring, while aggregation helps on seen questions but larger single models generalize better to unseen rubrics.

1 Introduction

Automated short answer grading (ASAG) can reduce assessment effort while maintaining consistent scoring quality (Burrows et al., 2015), but most resources target English. For German, only four dedicated ASAG corpora exist, with approximately 19,500 labeled answers compared to over 44,500 in English (Padó et al., 2024); morphology, compounding, case, and free word order further complicate transfer from English benchmarks (Horbach et al., 2024).

The BEA 2026 Shared Task 2 on Rubric-based Short Answer Scoring for German (Gombert et al., 2026) addresses this gap with the ALICE-LP-1.0 benchmark: 7,899 labeled student answers across 78 STEM questions, each paired with textual rubrics for **Correct**, **Partially correct**, and **Incorrect**. Because LLM grading accuracy degrades with rubric granularity (Deng et al., 2026) and ASAG transfer across datasets loses 13–47% QWK (Funayama et al., 2023), the task requires per-task adaptation rather than generic grading.

We present the WSE Research system (Figure 1), which combines three complementary strategies that address the following research questions:

RQ1 To what extent can rubric-conditioned prompt engineering close the performance gap between open-source and commercial LLMs for German short answer scoring?

RQ2 How does LoRA fine-tuning performance scale with model size for rubric-based short answer classification?

RQ3 Can ensemble aggregation over heterogeneous scoring strategies outperform individual models?

We evaluate three complementary design choices: rubric-conditioned prompting, LoRA fine-tuning, and aggregation across scorers with different error profiles. The strongest fine-tuned model surpasses the strongest prompted commercial baseline on the trial set (Ferreira Mello et al., 2025); on the official test set, the system places second on three tracks and third on the remaining one. Code and reproducibility artifacts are publicly available.[1]

2 Related Work

Automated Short Answer Grading. ASAG has progressed from keyword matching through supervised classifiers to neural approaches (Burrows et al., 2015); the SemEval-2013 winner already combined lexical, syntactic, and character *n*-gram features via stacking (Heilman and Mad-

*This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under grant FKZ 03FHP239.

[1] https://github.com/WSE-research/bea2026-german-asag

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1210–1216

nani, 2013). A recent survey reports that rubric-guided prompting and human-in-the-loop systems outperform zero-shot single-LLM setups (Nkoyo et al., 2025).

German ASAG. Padó et al. (2024) show that multilingual sentence embeddings are competitive on German but drop 15–20 F1 points on unseen questions, identifying cross-question generalization as the key bottleneck. Horbach et al. (2024) further show that multilingual transformers outperform translate-then-score pipelines across languages, motivating our German-language prompts and native fine-tuning.

LLM-Based Grading and Example Selection. LLMs support rubric-based scoring through in-context learning (Frohn et al., 2025), with rubric-conditioned prompts outperforming rubric-free approaches (Wang et al., 2019); yet fine-tuned smaller models can still outperform vanilla GPT-4 when labeled data is available (Ferreira Mello et al., 2025). Few-shot selection is similarly important: RAG-based retrieval outperforms random selection ($p < 0.001$) (Zhao et al., 2025), and boundary-focused exemplars reduce adjacent-score errors from 0.26 to 0.08 (Chu et al., 2026a). Our TF-IDF-based selection (Section 3.1) combines both signals. Prompt format is known to be brittle: rubric order and score notation can shift grades by up to 45% for frontier models (Deng et al., 2026); automated rubric refinement (Chu et al., 2026b) addresses this at the rubric side but was not attempted here.

Fine-Tuning and Calibration for ASAG. LoRA (Hu et al., 2022) enables parameter-efficient fine-tuning of large models on task-specific data. Funayama et al. (2023) report that rubric-keyphrase fine-tuning improves QWK by approximately 0.25 when training data is scarce. Raikote et al. (2026) reach QWK ≥ 0.80 with fine-tuned Qwen-7B on English short answers, while Bexte et al. (2024) show that confidence thresholds are prompt-dependent—a finding our experiments corroborate (Section 5.3). Our work extends LoRA fine-tuning to German with a systematic 7B–72B scaling study.

Taken together, prior work motivates our three-phase design. *Prompt engineering* operationalizes rubric-conditioned in-context scoring, *LoRA fine-tuning* tests whether task-specific supervision can replace commercial prompting, and a *retrieval baseline* anchors both against a non-neural alternative. We leave automated rubric refinement out of scope because the shared task evaluates systems on fixed organizer-provided rubrics; modifying them would confound scorer quality with rubric rewriting quality.

3 Approach

Our approach is designed as a three-phase rubric-scoring pipeline, as illustrated in Figure 1. First, we generate candidate score labels from the same input triple (question, student answer, rubric) with prompted LLMs, fine-tuned LLMs, and a TF-IDF kNN baseline that labels by similar training answers without neural inference. Second, we compare these scorers under trial-set evaluation to isolate the contribution of prompt engineering, model scale, and aggregation. Third, we choose the final prediction rule per deployment setting: aggregation for seen questions and a single large fine-tuned model for unseen questions, where rubric transfer dominates. The remainder of this section instantiates these phases through prompt engineering (Section 3.1), fine-tuning (Section 3.2), and prediction aggregation (Section 3.3).

3.1 Phase 1: Prompt Engineering

The prompt engineering phase instantiates the pipeline with rubric-conditioned in-context scoring. Each request includes the full textual rubric as the system prompt; the model receives the question and student answer and returns one of the three score labels. We evaluate this setup with Gemini 3 Flash and Qwen3.5-27B to compare a strong commercial API against the strongest self-hosted open-source prompting variant from our model-screening runs. This grounds scoring in the provided rubric rather than the model's parametric knowledge. We deliberately avoid chain-of-thought prompts, as recent evidence suggests CoT can degrade QWK on educational grading tasks despite improving NLG evaluation (Nkoyo et al., 2025).

TF-IDF-Based Example Selection. Rather than using fixed few-shot examples, we dynamically select training examples for each test instance via TF-IDF cosine similarity between the test answer and training answers for the same question. Examples enter the prompt in two roles: *similar examples*, retrieved by TF-IDF cosine similarity to the test answer among same-question training answers, and *boundary examples*, drawn from less-represented score labels to expose the model to

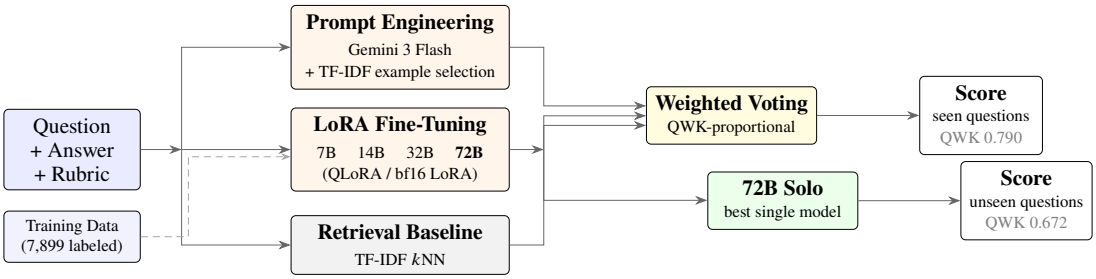

Figure 1: Overview of the WSE Research scoring pipeline. Three phases produce score predictions from the input triple (question, student answer, rubric). For seen questions, weighted voting is best; for unseen questions, the Qwen2.5-72B alone outperforms all ensembles. Dashed arrow: training data flow.

within-rubric edge cases. We quantify question difficulty as the dominant-label fraction in that question's training subset and bin questions into easy (> 0.60), medium (0.40–0.60), and hard (< 0.40) tiers; pilot sweeps showed two similar examples to be consistently helpful, while the useful number of boundary examples varied with difficulty. The final Gemini configuration, therefore, fixes the similar-example count at two and uses 1, 2, or 3 boundary examples for easy, medium, or hard questions, respectively. This combines retrieval of semantically close answers (Zhao et al., 2025) with boundary-focused calibration (Chu et al., 2026a). We iterated prompts systematically (6 strategies for Gemini 3 Flash, 33 variants across 8 rounds for Qwen3.5-27B); the checkpoint-specific transfer behavior this revealed is analyzed in Section 5.3.

3.2 Phase 2: Fine-Tuning

The fine-tuning phase asks how much rubric-conditioned supervision can be absorbed by open-source models of different sizes. We apply Low-Rank Adaptation (LoRA, cf. (Hu et al., 2022)) to Qwen-family models at four Qwen2.5 scales (Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, Qwen2.5-72B) plus Qwen3.5-9B to test whether a newer architecture improves fine-tuning efficiency at smaller scale. The Qwen3.5-27B model in Section 3.1 is used only for prompting, not for LoRA training.

Instruction Format. Each training example is formatted as an instruction-following turn: rubric text as system prompt, question and student answer as user message, gold score label as assistant response. This aligns fine-tuning with the inference-time interface from the prompt-engineering phase, so differences between prompting and fine-tuning can be attributed mainly to parameter adaptation rather than interface changes. Following Funayama et al. (2023), we include rubric key phrases as explicit input features.

Training Configuration. Preliminary studies using Qwen2.5-7B indicated that one epoch underfit and five epochs overfit, while increasing the LoRA rank beyond 32 did not improve trial performance. We therefore use a single pragmatic configuration across scales ($r = 32$, $\alpha = 32$, learning rate 2×10^{-4}, 3 epochs) to isolate model-size effects rather than retuning each model separately. The 14B, 32B, and 9B models use full bf16 LoRA on L40S GPUs; the Qwen2.5-7B and Qwen2.5-72B use 4-bit QLoRA (NF4), with the 72B moved to an H200 to fit memory constraints. We train each size in two configurations: *train-only* (for unbiased trial-set comparison) and *all-data* (train + trial, for final submission). Trial labels are released development data: train-only models provide an unbiased internal comparison, whereas all-data models use every labeled example available before hidden-test submission.

3.3 Phase 3: Prediction Aggregation

The aggregation phase asks whether complementary scorers should be combined or deployed separately. For unbiased trial-set analysis, we train a logistic-regression meta-learner (Heilman and Madnani, 2013) with leave-one-question-out cross-validation. Component models span three fine-tuned variants (Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B), the best prompt-only result (Qwen3.5-27B), and a non-neural TF-IDF k-nearest neighbor baseline. We set $k = 7$ because it was the strongest pure-retrieval setting in pilot runs, outperforming $k \in \{3, 5, 11\}$. Features comprise one-hot model predictions, kNN confidence, answer length, and question difficulty. For final submissions, we use simpler rules: QWK-weighted voting for seen questions and the 72B solo model for unseen questions, because weaker models diluted the 72B's rubric generalization.

4 Experimental Setup

Dataset. ALICE-LP-1.0 (Gombert et al., 2026) contains 7,899 German secondary-school STEM answers to 78 questions, each paired with a textual rubric defining three score levels. Train (7,072) and trial (827) sets are stratified by question (29% **Correct**, 36% **Partially correct**, 35% **Incorrect**). The test set (5,094 samples) splits into 2,008 answers to the 78 seen questions and 3,086 answers to 39 entirely new questions with 34 new rubrics. Each split is graded under both a 3-way label scheme and a 2-way scheme (with **Partially correct** collapsed into **Incorrect**), yielding the four official shared-task tracks.

Models and Hardware. We evaluate two prompt-only baselines (Gemini 3 Flash, parameter count undisclosed; Qwen3.5-27B) and five fine-tuned Qwen variants (Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, Qwen2.5-72B, Qwen3.5-9B), all trained with LoRA rank $r = 32$ for 3 epochs. Experiments up to 32B run on 2×L40S GPUs using vLLM (Kwon et al., 2023); the Qwen2.5-72B is fine-tuned on a single H200 NVL GPU via PEFT (Mangrulkar et al., 2023).

Evaluation. The primary metric is Quadratic Weighted Kappa (QWK), the official shared-task ranking metric; we additionally report accuracy. Trial-set evaluations use train-only models for unbiased comparison; submission models use all available data (train + trial).

5 Results

5.1 Trial Set Results

Table 1 reports trial-set performance with train-only fine-tuned models.

Fine-tuning consistently outperforms prompt engineering: the Qwen2.5-14B already surpasses Gemini 3 Flash (0.005 QWK), and the Qwen2.5-32B extends the margin to 0.021. A QWK-proportional weighted vote over the four-model pool {Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, Gemini 3 Flash}—with weights derived from each model's leave-one-question-out cross-validated QWK on the training set and normalized to sum to one—achieves the highest unbiased trial QWK of 0.781 by combining fine-tuned and prompt-based models; a logistic-regression stacker over an extended five-candidate pool that additionally includes the TF-IDF kNN baseline reaches 0.776 under leave-one-question-out cross-validation. This

Phase	System	QWK	Acc
Prompt Engineering			
	Gemini 3 Flash	0.748	73.6
	Qwen3.5-27B	0.719	70.2
Fine-Tuning (train-only → trial)			
	Qwen2.5-7B QLoRA	0.726	70.9
	Qwen2.5-14B LoRA	0.753	74.1
	Qwen2.5-32B LoRA	0.769	**75.7**
	Qwen2.5-72B QLoRA	0.768	**75.7**
	Qwen3.5-9B LoRA	0.756	74.2
Ensemble / Baseline			
	Weighted vote (4 models)	**0.781**	—†
	LogReg stacking (5-cand.)	0.776	75.5
	TF-IDF kNN ($k = 7$)	0.612	64.5

Table 1: Results on the ALICE-LP-1.0 trial set (827 samples). Fine-tuned models use train-only data for unbiased evaluation. † Accuracy is omitted for the QWK-proportional weighted vote because the aggregation produces a continuous score in $[0, 2]$ rather than a discrete label, and any threshold-based discretization is a separate decision rule unrelated to the QWK ranking metric.

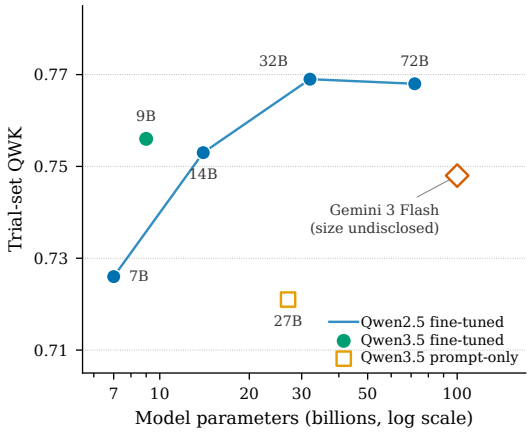

Figure 2: Trial-set QWK versus model size: fine-tuned Qwen variants (filled) saturate by 32B; prompt-only baselines (hollow); Gemini 3 Flash is plotted at a nominal 100B (true size undisclosed).

answers RQ1: rubric-conditioned prompt engineering yields a strong baseline, but it does not close the gap to fine-tuned open-source models once task-specific supervision is available.

Scaling Behavior. Figure 2 visualizes trial-set QWK as a function of model size. Fine-tuning QWK scales with diminishing returns and saturates by 32B (−0.001 from 32B to 72B). The 7B (0.726) already matches the prompt-only Qwen3.5-27B (0.719); Qwen3.5-9B (0.756) outperforms Qwen2.5-14B (0.753), suggesting that model generation matters more than parameter count. This answers RQ2: larger fine-tuned models generalize more robustly across rubrics, but

Track	Ours	IWM-DKM	Rank
3-way scoring			
Seen questions	**0.790**	0.796	**2nd**
Unseen questions	**0.672**	0.681	**2nd**
2-way scoring			
Seen questions	**0.717**	0.726	**2nd**
Unseen questions	**0.533**	0.550	**3rd**

Table 2: Test-set QWK on all four official tracks for our system and the winning team, IWM-DKM. Our submitted systems are the QWK-weighted vote of {Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, Gemini 3 Flash} on 3-way seen, the binary majority vote of {Qwen2.5-14B, Qwen2.5-32B, Gemini 3 Flash} on 2-way seen, and the Qwen2.5-72B solo on both unseen-question tracks. Retraining the 72B on train + trial lifts its trial QWK from 0.768 (train-only, Table 1) to 0.851 (in-sample), a 0.083 headroom from incorporating trial labels at submission time.

unbiased trial performance saturates by 32B.

5.2 Test Set Results

Table 2 reports our best system on each of the four official tracks alongside the corresponding result of the winning team IWM-DKM.

On *seen questions*, the weighted ensemble (QWK 0.790) outperforms all individual models, mirroring the trial-set pattern. On *unseen questions*, the Qwen2.5-72B alone (QWK 0.672) beats a 72B-inclusive weighted vote (0.653) and the Qwen2.5-32B solo (0.611); adding weaker models *hurts*, and the seen-to-unseen drop is 0.118 QWK for 72B versus 0.169 for 32B. This answers RQ3: aggregation helps on seen questions, but becomes harmful on unseen ones once a single strong model dominates.

5.3 Analysis

Prompt-Model Coupling and Calibration. Optimal prompts are checkpoint-specific: prompts tuned for Gemini 3 Flash degrade by 0.260–0.400 QWK on other Gemini variants, aligning with Deng et al. (2026) on format sensitivity. Fine-tuned models exhibit >99% logprob-confidence, making confidence-based routing infeasible without post-hoc calibration (Raikote et al., 2026).

Error Patterns. On the trial set (32B train-only), 94.3% of misclassifications are adjacent-boundary errors, and **Partially correct** has the lowest per-label precision (65.0%), consistent with rubric-granularity findings (Deng et al., 2026). Cross-model agreement drops from 53.9% on seen to 41.4% on unseen, where the Qwen2.5-72B most often corrects majority leniency by downgrading.

6 Conclusion

We presented the WSE Research system for the BEA 2026 Shared Task 2, placing second on three of four tracks. Rubric conditioning matters, parameter-efficient fine-tuning exploits it most reliably and saturates by 32B (RQ1, RQ2), and aggregation helps on seen questions but hurts unseen ones (RQ3). Rubric-conditioned fine-tuning is thus a practical open-source alternative.

Limitations

Our system was developed and evaluated exclusively on the ALICE-LP-1.0 dataset, which covers German STEM answers at the secondary-school level. Generalization to other languages, educational levels, or subject domains has not been tested. Prior work suggests that ASAG strategies degrade by 13–47% QWK when transferred across datasets (Funayama et al., 2023), and cross-corpus transfer for German specifically can incur up to 22 percentage points of accuracy loss (Padó et al., 2024).

The 72B out-of-sample trial QWK (0.768) shows no improvement over the 32B (0.769), suggesting that the scaling curve plateaus at 32B for this dataset size (7,072 training samples). Retraining the 72B on train + trial for submission lifts its trial QWK to 0.851 (in-sample), a 0.083 headroom over the directly comparable train-only value; this gain is observed on data the model has seen and should not be read as out-of-sample improvement.

We rely on the provided rubrics as-is and do not perform rubric quality assessment or optimization. Approximately 10% of questions use vague competency-level descriptions without specific criteria, which may limit scoring accuracy for both our system and human raters. Confusion-aware rubric refinement (Chu et al., 2026b) could address this but was not attempted in the current work.

All evaluations use automated metrics (Quadratic Weighted Kappa, accuracy); we did not conduct human evaluation of model predictions or analyze inter-annotator agreement on the dataset.

Fine-tuned models show near-uniform confidence (>99%), which prevents meaningful confidence-based quality control. This limits the applicability of human-in-the-loop strategies that rely on uncertainty estimates for selective routing (Bexte et al., 2024). Post-hoc temperature scaling (Raikote et al., 2026) may restore useful confidence

signals but was not applied here.

Ethics Statement

This work uses the publicly released ALICE-LP-1.0 dataset, which was collected with appropriate institutional oversight within the ALICE project (Leibniz Foundation). Student answers are anonymized and contain no personally identifiable information.

Automated scoring systems should be deployed as decision-support tools rather than sole arbiters of student grades. We note that LLM-based scoring can reproduce or amplify biases present in training data, including potential biases in rubric design or human scoring patterns. Deployment in high-stakes assessment contexts requires human oversight and ongoing bias monitoring.

Generative AI Disclosure. Parts of the codebase were developed with the assistance of AI agents. All experimental design, analysis, and writing were conducted by the authors.

References

Marie Bexte, Andrea Horbach, Lena Schützler, Oliver Christ, and Torsten Zesch. 2024. Scoring with confidence? – Exploring high-confidence scoring for saving manual grading effort. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 119–124. Association for Computational Linguistics.

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(1):60–117.

Yucheng Chu, Hang Li, Kaiqi Yang, Yasemin Copur-Gencturk, Kevin Haudek, Joseph Krajcik, and Jiliang Tang. 2026a. Optimizing in-context demonstrations for LLM-based automated grading. *Preprint*, arXiv:2603.00465.

Yucheng Chu, Hang Li, Kaiqi Yang, Yasemin Copur-Gencturk, Joseph Krajcik, Namsoo Shin, and Jiliang Tang. 2026b. Confusion-aware rubric optimization for LLM-based automated grading. *Preprint*, arXiv:2603.00451.

Haotian Deng, Chris Farber, Jiyoon Lee, and David Tang. 2026. Rubric-conditioned LLM grading: Alignment, uncertainty, and robustness. *Preprint*, arXiv:2601.08843.

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 2025)*, pages 93–103.

Scott Frohn, Tyler J. Burleigh, and Jing Chen. 2025. Automated scoring of short answer questions with large language models: Impacts of model, item, and rubric design. In *Artificial Intelligence in Education (AIED 2025)*, volume 15882 of *Lecture Notes in Artificial Intelligence*, pages 44–51. Springer.

Hiroaki Funayama, Yuichiroh Shindo, Hiroki Ouchi, and Kentaro Inui. 2023. Reducing the cost: Cross-prompt pre-finetuning for short answer scoring. In *Proceedings of the 24th International Conference on Artificial Intelligence in Education (AIED 2023)*.

Sebastian Gombert, Zhifan Sun, Jannik Lossjew, Tobias Wyrwich, Berrit Katharina Czinczel, David Bednorz, Sascha Bernholt, Knut Neumann, Ute Harms, Aiso Heinze, Fabian Zehner, and Hendrik Drachsler. 2026. BEA 2026 shared task: Rubric-based short answer scoring for german. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. To appear. Task page: `https://edutec.science/bea-2026-shared-task/`.

Michael Heilman and Nitin Madnani. 2013. ETS: Domain adaptation and stacking for short answer scoring. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 275–279.

Andrea Horbach, Joey Pehlke, Ronja Laarmann-Quante, and Yuning Ding. 2024. Crosslingual content scoring in five languages using machine-translation and multilingual transformer models. *International Journal of Artificial Intelligence in Education*, 34(4):1294–1320.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP 2023)*, pages 611–626.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2023. PEFT: Parameter-efficient fine-tuning of billion-scale models on low-resource hardware.

Fredrick Eneye Tania-Amanda Nkoyo, Chukwuebuka Fortunate Ijezue, Maaz Amjad, Ahmad Imam Amjad, Sabur Butt, and Gerardo Castaneda-Garza. 2025. Advances in auto-grading with large language

models: A cross-disciplinary survey. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*.

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

Pranav Raikote, Korbinian Randl, Ioanna Miliou, Athanasios Lakes, and Panagiotis Papapetrou. 2026. CHiL(L)Grader: Calibrated human-in-the-loop short-answer grading. *Preprint*, arXiv:2603.11957.

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.

Chenyan Zhao, Mariana Silva, and Seth Poulsen. 2025. Language models are few-shot graders. In *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 3–16. Springer Nature Switzerland.

A Prompt Templates

This appendix gives the verbatim wording of the two prompt families used in the prompt-engineering phase (Section 3.1) and as the inference-time interface for the fine-tuning phase (Section 3.2).

Prompted scoring (Gemini 3 Flash, Strategy C5c). The user prompt contains the question, the verbatim three-level rubric, two retrieved similar examples, $n_b \in \{1, 2, 3\}$ boundary examples (selected as described in Section 3.1), and finally the test answer. The system prompt is verbatim:

```
Du bist ein automatisches Bewertungs-
system für Schülerantworten in
MINT-Fächern.
AUFGABE
Bewerte die Antwort eines Schülers
anhand der Bewertungsrubrik und der
Beispielbewertungen.
BEWERTUNGSSTUFEN
- ‘‘Correct’’: Die Antwort
trifft ALLE zentralen Punkte der
Correct-Rubrik.
- ‘‘Partially correct’’: Die Antwort
trifft MINDESTENS EINEN zentralen
fachlichen Punkt der Rubrik korrekt,
verfehlt aber andere wesentliche
Kriterien.
- ‘‘Incorrect’’: Die Antwort trifft
KEINEN der fachlichen Kernpunkte der
Rubrik.
ENTSCHEIDUNGSREGELN
1. Entscheide anhand der
Rubrik UND konsistent mit den
Beispielbewertungen.
2. Bewerte nur, was geschrieben
wurde --- keine wohlwollenden
Annahmen.
3. GRENZE Incorrect vs. Partially
correct:
  - Vage Aussagen, Alltagswissen
oder Umformulierungen der Frage OHNE
fachlichen Inhalt aus der Rubrik sind
IMMER Incorrect.
  - Partially correct erfordert
NACHWEISBAR mindestens einen
konkreten fachlichen Punkt, der
in der Rubrik als Kriterium genannt
wird.
4. GRENZE Partially correct vs.
Correct:
  - Correct erfordert, dass die
Antwort die Rubrik-Kriterien
VOLLSTÄNDIG abdeckt.
  - Wenn die wesentlichen
Konzepte korrekt und vollständig
dargelegt sind, kleine sprachliche
Ungenauigkeiten aber vorliegen ->
trotzdem Correct.
  - Im Zweifel: orientiere dich an
den Beispielbewertungen für diese
Frage.
5. Leere Antworten, Nichtwissen
(‘?’, ‘Keine Ahnung’), einzelne
Wörter ohne Erklärung -> IMMER
Incorrect.
FORMAT
Antworte NUR mit einem JSON-Objekt:
{‘‘score’’: ‘‘Correct’’
| ‘‘Partially correct’’ |
‘‘Incorrect’’, ‘‘confidence’’:
0.0--1.0}
```

Fine-tuned scoring (Qwen2.5 LoRA). The fine-tuned models receive a minimal instruction-format prompt that mirrors the inference contract and dispenses with explicit decision rules, since those have been internalized via training:

```
System: Du bist ein Bewertungssystem
für Schülerantworten. Bewerte die
Antwort anhand der Rubrik. Antworte
ausschliesslich mit JSON: {‘‘score’’:
‘‘Correct’’ | ‘‘Partially correct’’ |
‘‘Incorrect’’}
User: Frage: {question} \n
Bewertungsrubrik: - Correct:
{rubric.Correct} - Partially
correct: {rubric.PC} - Incorrect:
{rubric.Inc} \n Schülerantwort:
{answer}
```

The full executable prompt code, including example selection and all model-specific variants, is publicly available in our repository (see Section 1).