WSE-Research at CLEF 2026 ELOQUENT Sensemaking: LLM-Based Scoring of Student Answers to PISA Reading Comprehension Items

ELOQUENT Lab at CLEF 2026

Jonas Gwozdz[1,*], Andreas Both[1]

[1]Leipzig University of Applied Sciences (HTWK Leipzig), Faculty of Computer Science and Media, Leipzig, Germany

Abstract

This paper describes the WSE-Research system submitted to the CLEF 2026 ELOQUENT Sensemaking shared task, which evaluates automated scoring of student answers to PISA reading comprehension items across 13 languages. Our system uses a single open-weights language model (Qwen3.5-27B) with a track-conditional prompting strategy: few-shot retrieval-augmented prompting for the rubric-based track and zero-shot prompting for the simple rating track, combined with constrained decoding to ensure valid output format. The system handles all 13 languages natively without translation or language-specific components. Our approach ranked first on both tracks, achieving a quadratic weighted kappa (QWK) of 0.65 on the rubric-based track and 0.46 on the simple rating track, placing first on every individual language in both tracks.

Keywords

automated short answer grading, large language models, PISA, multilingual assessment, CLEF ELOQUENT, constrained decoding

1. Introduction

The Programme for International Student Assessment (PISA), administered by the Organisation for Economic Co-operation and Development (OECD), evaluates reading, mathematics, and science competencies of 15-year-old students across more than 80 participating countries. As PISA moves toward digital delivery and open-response items, the demand for reliable automated scoring of student answers grows. Automated Short Answer Grading (ASAG) has progressed from feature-engineered classifiers [1] to large language model (LLM)-based approaches that can score student responses without task-specific training data [2].

The ELOQUENT Sensemaking shared task at CLEF 2026 [3] provides a standardized benchmark for automated scoring of PISA-style reading comprehension answers. The task spans two evaluation tracks—**Rubric**-based rating (3-way classification) and **Simple** rating (5-point ordinal scale)—across 13 languages, making it one of the most linguistically diverse ASAG benchmarks to date. Participant systems must score student answers grounded in reading passages of 1,200 to 4,300 characters, a setting that tests both language understanding and rubric interpretation.

In this paper, we describe the WSE-Research system that ranked first on both tracks. Our approach uses a single 27-billion-parameter open-weights model (Qwen3.5-27B [4, 5]) served via vLLM [6] on a single GPU. We employ a track-conditional prompting strategy: few-shot retrieval-augmented prompting for the **Rubric** track, where retrieved examples calibrate the model's scoring, and zero-shot prompting for the **Simple** track, where few-shot examples were found to degrade performance during development. Constrained decoding ensures that every model output is a valid label, eliminating parse failures entirely. The system handles all 13 languages natively without translation or language-specific preprocessing.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

jonas.gwozdz@htwk-leipzig.de (J. Gwozdz); andreas.both@htwk-leipzig.de (A. Both)

0009-0007-8348-2357 (J. Gwozdz); 0000-0002-9177-5463 (A. Both)

The remainder of this paper is organized as follows. Section 2 reviews related work on automated short answer grading. Section 3 describes the shared task, data, and evaluation methodology. Section 4 presents our system architecture and design decisions. Section 5 reports the official results. Section 6 analyzes performance patterns across languages and data conditions. Section 7 concludes with limitations and future directions.

2. Related Work

Automated Short Answer Grading has evolved substantially over the past decade. Burrows *et al.* [1] provide a comprehensive survey covering three eras of ASAG: hand-crafted rules, statistical methods, and early neural approaches. More recently, Ferreira Mello *et al.* [2] examine the impact of LLMs on ASAG, finding that zero-shot and few-shot prompting of instruction-tuned models can match or exceed supervised baselines on standard benchmarks. Frohn *et al.* [7] further show that rubric detail is a stronger moderator of LLM scoring quality than model choice, motivating our use of the full rubric text in the prompt.

Few-shot example retrieval has emerged as a key technique for LLM-based grading. Zhao *et al.* [8] demonstrate that retrieval-augmented few-shot prompting significantly outperforms random example selection for ASAG ($p < 0.001$), validating the TF-IDF retrieval approach we adopt. In a closely related shared task, we achieved second place at the BEA 2026 Shared Task on Rubric-Based Short Answer Scoring for German [9], using a multi-strategy pipeline with weighted voting across model configurations. The present work adapts and simplifies that approach for the multilingual PISA setting.

Multilingual automated scoring remains an active challenge. Horbach *et al.* [10] evaluate cross-lingual content scoring in five languages, finding that multilingual transformer models outperform translate-then-score pipelines. For PISA specifically, Zehner *et al.* [11] describe *Omethi*, a hierarchical cascade system that scored 3.8 million PISA responses across 11 languages, achieving $\kappa = 0.804$. Our work differs from *Omethi* in using zero-shot LLM prompting rather than fine-tuned encoder cascades, and in handling 13 languages with a single model.

The Sensemaking shared task has been part of the ELOQUENT lab at CLEF since 2025. Šindelář and Bojar [12] provide an overview of the 2025 edition, establishing the task format and evaluation methodology. The 2026 edition is described in the ELOQUENT lab overview [13] and the accompanying task overview [3]. Šindelář *et al.* [14] describe the generation of training data for context-dependent rubric-based grading, including the synthetic data pipeline used to create parts of the 2026 shared task data.

3. Task Description

3.1. Overview

The ELOQUENT Sensemaking shared task at CLEF 2026 [3] benchmarks automated scoring of student answers to reading comprehension questions grounded in longer reading passages. Items cover diverse domains including popular science, world history, legal texts, and counseling. The task spans 13 languages: English, Czech, Portuguese, German, Hungarian, Finnish, Danish, Swedish, Serbian, Greek, Irish, Romanian, and Ukrainian. Two independent evaluation tracks are offered, each scored on the full multilingual test set.

3.2. Tracks

Rubric-Based Rating. Given a quadruple of *(context, question, rubric, answer)*, the system assigns one of three labels: Full Credit (2), Partial Credit (1), or No Credit (0). The rubric provides prose descriptions defining the criteria for each label, always in English regardless of the item language.

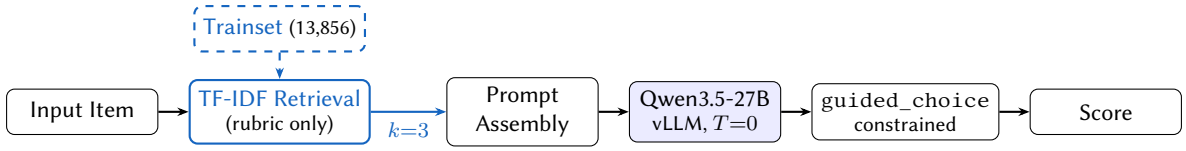

Figure 1: System overview. For the **Rubric** track, three similar examples are retrieved from the training set via TF-IDF similarity (blue path). For the **Simple** track, no retrieval is performed. Constrained decoding ensures every output is a valid label.

Simple Rating. Given a triple of *(context, question, answer)* without a rubric, the system assigns a score on a 0–4 ordinal scale, where 0 indicates no relevant effort and 4 indicates a fully correct answer.

3.3. Data

The development set contains 4,146 items distributed approximately evenly across the 13 languages (∼320 items per language). The training set provides 13,856 multilingual labeled items and an additional 18,377 English-only items, covering domains that partially overlap with the development set. The test set is substantially larger, comprising approximately 35,000 PISA items (under OECD NDA) and approximately 47,000 non-PISA items per track.

The development set label distribution for the rubric track is 28% No Credit, 28% Partial Credit, and 44% Full Credit. For the simple track, labels are skewed: label 4 (fully correct) accounts for 43% of items, while label 3 comprises only 6%.

Items vary in origin: contexts are predominantly auto-translated from English, rubrics are largely LLM-generated with human correction, and answers include both human-written and LLM-generated responses [14].

3.4. Evaluation

Systems are evaluated using Quadratic Weighted Kappa (QWK), computed per domain following the task definition, where a *domain* denotes a distinct thematic part of the test set (the non-PISA data alone spans several domains). The final ranking is the equal-weight mean of QWK over the two portions of the test set (non-PISA and PISA), giving both the same weight regardless of item count.

4. System Description

Figure 1 provides an overview of our scoring pipeline. The system processes each item independently: an input item is assembled into a prompt (with optional few-shot examples for the rubric track), scored by Qwen3.5-27B, and decoded into a valid label via constrained decoding.

4.1. Model

We use Qwen3.5-27B [4, 5],[1] an instruction-tuned open-weights language model with 27 billion parameters. The checkpoint corresponds to the model's Hugging Face `main` revision retrieved on 2026-04-14; a specific commit revision was not pinned at inference time. The model is served in native `bfloat16` without quantization via vLLM [6] (version 0.19.0) on a single NVIDIA H200 GPU with 141.00 GB VRAM at HTWK Leipzig. We set the temperature to 0.0 for deterministic inference and disable the model's reasoning mode to ensure direct label output. Inference uses a context window of 16,384 tokens; the longest inputs (a reading passage of up to 4,300 characters plus the rubric and, on the rubric track, three retrieved examples) stay well within this window, so no item required truncation. Scoring throughput is approximately 1.8 items per second in zero-shot mode and 1.4 items per second with few-shot retrieval (longer prompts).

[1] https://huggingface.co/Qwen/Qwen3.5-27B

4.2. Prompting Strategy

We employ a *track-conditional* prompting strategy, using different approaches for each track based on development set analysis.

Rubric-Based Track: Few-Shot Prompting. The prompt consists of a system message instructing the model to score based only on the rubric criteria, followed by the reading passage, question, rubric descriptions for each label (Full Credit, Partial Credit, No Credit), three retrieved few-shot examples with their gold labels, and the student answer to score. The model is constrained to output exactly one of: 0, 1, or 2.

Simple Rating Track: Zero-Shot Prompting. The prompt contains a system message instructing the model to rate answer quality on a 0–4 scale, followed by the reading passage, question, short descriptions of each score level, and the student answer. No few-shot examples are included.

Rationale. The decision to use different strategies per track is empirically motivated. On the development set, few-shot retrieval improved rubric scoring by a modest $+0.01$ QWK but degraded simple rating by -0.03 QWK. While the rubric improvement is small in absolute terms, the direction was consistent across cross-validation folds and languages, and the asymmetry between tracks was large enough to motivate a track-conditional design. We hypothesize that the three rubric labels are well-defined by the rubric text, making retrieved examples useful calibration anchors, whereas the five simple rating labels are more subjective, and retrieved examples over-calibrate the model toward item-specific patterns.

4.3. Few-Shot Retrieval

For the rubric-based track, we retrieve three training examples per test item using TF-IDF similarity over answer texts. The TF-IDF index is built over all 13,856 multilingual training set answers with 20,000 features, bigram tokenization, and sublinear term frequency weighting. For each test item, the three most similar training items by cosine similarity of their answer-text TF-IDF vectors are retrieved and inserted into the prompt between the rubric description and the target answer. Index construction takes approximately two seconds; per-item retrieval is negligible.

4.4. Constrained Decoding

We use vLLM's `guided_choice` parameter to constrain the model's output to the valid label set for each track: {0, 1, 2} for rubric-based rating and {0, 1, 2, 3, 4} for simple rating. This standard decoding feature eliminates parse failures, ensuring 100% format compliance across all approximately 170,000 scored items without output parsing or retry logic.

4.5. Multilingual Handling

All 13 languages are handled natively by Qwen3.5-27B without translation. The system prompt and score descriptions are in English; the reading passage, question, and student answer remain in the target language as provided. No language detection, language-specific routing, or preprocessing is applied.

4.6. Submission

The shared task allows up to five runs per team. We originally planned a diverse set of runs including stacking ensembles that achieved QWK 0.87 on the development set rubric track. However, the test set was approximately 8.4 times larger than anticipated ($\sim$82,000 items per track vs. 4,146 in the development set), making the full ensemble pipeline infeasible within the submission deadline. We submitted two runs: (1) our PRIMARY submission described above (Qwen3.5-27B, track-conditional),

Table 1
Official rankings. **Left: Rubric**-based rating track. **Right: Simple** rating track. Participant teams in **bold**; organizer baselines in regular type. Rubric baselines with QWK < 0.25 omitted.

#	System	n-P	PISA	All
1	**WSE (ours)**	**.713**	**.583**	**.648**
2	Gemma-4-E4B	.536	.475	.506
3	**fianso**	.476	.391	.434
4	EuroBERT-210m	.539	.123	.331
5	**pythoneers**	.449	.185	.317
6	Gemma-4-E2B	.317	.255	.286
7	**writerslogic**	.433	.053	.243
8	**dipf_tba**	.226	.149	.188

#	System	n-P	PISA	All
1	**WSE (ours)**	**.416**	**.512**	**.464**
2	Gemma-4-E4B	.299	.456	.378
3	Gemma-4-E2B	.256	.273	.265
4	**deep_thinkers**	$-.018$	$-.043$	$-.031$

and (2) a model family hedge using Gemma-3-27B [15] with zero-shot prompting on both tracks. Both submissions use deterministic decoding ($T{=}0$) with the model's reasoning mode disabled. Although Qwen3.5-27B is reasoning-capable and the task provided a `misc` field for recording chain-of-thought traces, we submitted a single deterministic run per system and did not emit reasoning as a secondary output; with $T{=}0$ scoring, repeated runs would be redundant.

5. Results

5.1. Official Rankings

Table 1 shows the official final rankings for both tracks (left: rubric-based rating; right: simple rating). WSE-Research ranked first on both tracks by a substantial margin.

On the rubric-based track (Table 1, left), our system achieves an overall QWK of 0.65, with 0.71 on the non-PISA test set and 0.58 on the PISA test set. The next-best participant team (fianso) scores 0.43, a gap of $+0.21$. The strongest organizer baseline (Gemma-4-E4B) scores 0.51.

Our secondary submission (Gemma-3-27B, zero-shot on both tracks) achieved PISA rubric QWK 0.46, confirming that the approach is not specific to the Qwen model family.

On the simple rating track (Table 1, right), our system achieves an overall QWK of 0.46. Only two participant teams submitted to the simple track; the strongest organizer baseline scores 0.38.

5.2. Per-Language Performance

Table 2 shows QWK for each of the 13 languages on the combined test set (PISA + non-PISA). Our system ranked first on every individual language in both tracks.

Rubric QWK ranges from 0.50 (Ukrainian) to 0.74 (Romanian), a spread of 0.24. Simple QWK is more compressed, ranging from 0.32 (Ukrainian) to 0.47 (Portuguese), suggesting that the task's inherent difficulty dominates language-specific effects on this track.

6. Analysis

6.1. Error Patterns

To understand where scoring errors concentrate, we examine confusion matrices on the development set (4,146 items) using our zero-shot Qwen3.5-27B predictions.

Rubric track. Table 2 (right) shows the rubric confusion matrix. No Credit (0) and Full Credit (2) are scored reliably (recall 0.94 and 0.85, respectively). Partial Credit (1) is the systematic weak point: only 0.61 recall, with 31% of Partial Credit answers misclassified as No Credit. The model exhibits a conservative bias, tending to under-score rather than over-score partial answers.

Table 2
Left: Per-language QWK on the combined test set (WSE-Research, ranked 1st on every language). **Right:** **Rubric** track confusion matrix on the development set (zero-shot). Partial Credit is the weakest label (recall 0.61).

	Rubric		Simple	
Lang.	QWK	Acc	QWK	Acc
ro	.736	.671	.426	.314
pt	.666	.711	**.473**	.338
el	.662	.699	.434	.311
sr	.658	.697	.438	.315
sv	.650	.695	.435	.310
ga	.647	.674	.405	.289
de	.633	.684	.472	.325
hu	.632	.682	.434	.300
da	.625	.679	.433	.304
cs	.621	.667	.409	.310
en	.615	.660	.423	.318
fi	.568	.651	.428	.312
uk	.497	.604	.317	.270

Romance Germanic Slavic Uralic Celtic Hellenic

	Pred NC	Pred PC	Pred FC	Recall
Gold NC (0)	**1086**	66	2	.94
Gold PC (1)	365	709	91	.61
Gold FC (2)	11	263	**1553**	.85

Acc: 0.81, $N = 4{,}146$

31% of Partial Credit items are misclassified as No Credit. The model under-scores rather than over-scores.

Table 3
Per-language QWK on non-PISA vs. PISA test data. Romanian has no PISA items. On rubric, PISA is consistently harder. On simple, the pattern reverses: PISA is easier for 11 of 12 languages.

	Rubric-Based			Simple Rating		
	non-P	PISA	Δ	non-P	PISA	Δ
cs	.711	.583	−.13	.383	.559	+.18
da	.728	.577	−.15	.419	.524	+.11
de	.747	.577	−.17	.460	.556	+.10
el	.723	.636	−.09	.418	.529	+.11
en	.629	.588	−.04	.416	.494	+.08
fi	.717	.497	−.22	.413	.501	+.09
ga	.732	.565	−.17	.397	.524	+.13
hu	.714	.598	−.12	.407	.600	+.19
pt	.765	.614	−.15	.465	.527	+.06
ro	.736	—	—	.426	—	—
sr	.726	.626	−.10	.431	.483	+.05
sv	.727	.617	−.11	.422	.528	+.11
uk	.676	**.080**	−.60	.337	.113	−.22

Simple track. On the 5-point scale, the model achieves reasonable recall only for label 4 (fully correct, recall 0.68). Labels 0–2 cluster around 0.40 recall with heavy cross-confusion, and label 3 is nearly unpredictable (recall 0.22), consistent with label 3 comprising only 6% of the development set. The difficulty of distinguishing adjacent ordinal categories confirms the track comparison finding: the rubric track's coarser label space is inherently more tractable.

6.2. PISA vs. Non-PISA by Language

Table 3 disaggregates performance by PISA vs. non-PISA data for each language. Two patterns stand out.

Ukrainian PISA failure. Ukrainian PISA rubric QWK is 0.08—effectively random for 3-class classification—while non-PISA Ukrainian reaches a reasonable 0.68. The combined score of 0.50

reported in Table 2 masks this catastrophic PISA failure. Ukrainian is also the only language where PISA simple performance (0.11) is worse than non-PISA (0.34). The cause is unclear: it may reflect specific properties of the Ukrainian PISA items, Cyrillic-script weaknesses in the model, or an interaction of both. This result should not be over-interpreted: the Ukrainian PISA section is considerably smaller than the other languages' and lacks answers for some questions, so its QWK estimate is correspondingly less reliable.

Reversed simple track pattern. On the **Rubric** track, **PISA** data is harder for every language (all deltas negative). On the **Simple** track, the pattern reverses: **PISA** QWK exceeds **non-PISA** QWK for 11 of 12 languages, with Hungarian showing the largest gain (+0.19). This suggests that the PISA simple rating labels may be more consistently annotated than the non-PISA labels, or that real student answers are paradoxically easier to rank on a coarse ordinal scale than LLM-generated answers.

Romanian. Romanian has no PISA items, which partly explains its top-ranking rubric QWK of 0.74: it is evaluated only on the easier non-PISA data.

6.3. Track Comparison

The rubric-based track yields substantially higher QWK (0.65) than the simple rating track (0.46) across all systems. This reflects both the smaller label space (3 vs. 5 classes) and the availability of rubric text as a scoring anchor. For reference, the *Omethi* system [11] achieved $\kappa = 0.804$ on PISA rubric-based scoring with fine-tuned encoder cascades—substantially above our zero-shot result of 0.58 on PISA data, indicating that the operational scoring ceiling remains distant.

6.4. Development Set to Test Set Generalization

Table 4 compares all scoring approaches we evaluated on the development set. Among single open-weights models, Qwen3.5-27B was the strongest on both tracks, which is why the submitted system is built on it. Adding few-shot retrieval helped the rubric track but hurt the simple track, motivating the track-conditional design (few-shot for the rubric track, zero-shot for the simple track). Stacking ensembles scored highest on the development set, but the expanded stacker requires running nine models and was infeasible on the far larger test set, so it was never submitted.

Our development set performance was substantially higher than test set performance: rubric QWK 0.87 vs. 0.65 overall, and simple QWK 0.71 vs. 0.46. This drop is consistent across all participating systems and reflects domain shift, PISA item difficulty, and different label distributions.

Despite this gap, our system maintained its first-place ranking, suggesting that the zero-shot approach generalizes better to unseen domains than more domain-dependent methods. Fine-tuned encoder models showed particularly large drops on the PISA subset (e.g., EuroBERT-210m: non-PISA 0.54 vs. PISA 0.12).

7. Conclusion

We described the WSE-Research system for the CLEF 2026 ELOQUENT Sensemaking shared task. Using a single open-weights model (Qwen3.5-27B) with track-conditional prompting and constrained decoding, our system ranked first on both evaluation tracks across all 13 languages. The approach requires no fine-tuning, no translation pipeline, and no language-specific components.

Our analysis reveals that Partial Credit is the hardest label for LLM-based grading (31% misclassified as No Credit), and that the PISA/non-PISA performance gap varies dramatically by language—Ukrainian PISA rubric scoring is essentially random (QWK 0.08), while non-PISA Ukrainian reaches 0.68. We also find that few-shot retrieval helps rubric-based scoring but degrades rubric-free scoring, motivating our track-conditional strategy.

Table 4
Development-set QWK (4,146 items) for every scoring approach we evaluated. Single models and the voting ensemble are scored on the full development set; stackers use deterministic 5-fold cross-validation. Few-shot uses three TF-IDF-retrieved examples. The voting ensemble is track-conditional: majority vote over the top-5 models on the rubric track and ordinal mean over the top-3 models on the simple track. The submitted PRIMARY run is track-conditional, taking the few-shot rubric result (0.866) and the zero-shot simple result (0.711) of Qwen3.5-27B. The expanded stacker scored highest but was infeasible on the far larger test set.

Approach	Rubric	Simple
Single open-weights models (zero-shot)		
Qwen3.5-27B	.861	.711
Qwen3.6-27B	.858	.641
Gemma-3-27B	.852	.699
Qwen2.5-32B	.844	.685
Qwen3-32B	.839	.623
Llama-3.1-70B (AWQ)	.834	.705
Qwen2.5-72B (AWQ)	.830	.657
Qwen3.5-9B	.823	.642
Qwen2.5-7B	.755	.474
Qwen3.5-27B with few-shot retrieval		
Few-shot (k=3)	.866	.679
Ensembles and stackers		
Voting ensemble	.858	.745
Stacker (5 zero-shot)	.866	.841
Stacker (5 zero-shot + 4 few-shot)	**.873**	**.847**
Organizer baseline	.760	.360

The main limitations of this work are: (1) our results (rubric QWK 0.65) remain below the operational thresholds typically required for large-scale assessment deployment [11]; (2) Ukrainian PISA scoring is essentially random (though this section is small and incomplete, so it should not be over-interpreted), and the cross-linguistic performance variance (QWK range 0.08–0.74 on PISA rubric) raises fairness concerns for any deployment scenario; (3) the stacking ensemble that achieved our best development set results could not be evaluated on the test set due to its unexpected scale; and (4) no significance tests are reported due to test set gold labels being under NDA.

Future work will explore ensemble methods at scale, language-aware few-shot retrieval, and domain adaptation for PISA items. We plan to test whether translation-based scoring can recover Ukrainian performance, and whether the track-conditional pattern generalizes to other grading tasks and label granularities.

Acknowledgments

We thank the OECD for providing access to the PISA data under NDA, Pavel Šindelář and Ondřej Bojar at Charles University for organizing the Sensemaking shared task, and HTWK Leipzig for providing GPU infrastructure (NVIDIA H200). This work was conducted as part of the first author's doctoral research at HTWK Leipzig, with funding from the German Federal Ministry of Education and Research (BMBF, Förderkennzeichen 03FHP239).

Declaration on Generative AI

Claude (Anthropic) was used during manuscript preparation for editorial review, section drafting, and code generation for the scoring pipeline. All AI-generated content was reviewed, verified, and revised by the authors, who take full responsibility for the final manuscript.

References

[1] S. Burrows, I. Gurevych, B. Stein, The eras and trends of automatic short answer grading, International Journal of Artificial Intelligence in Education 25 (2015) 60–117. doi:10.1007/s40593-014-0026-8.

[2] R. Ferreira Mello, C. Pereira Junior, L. Rodrigues, F. D. Pereira, L. Cabral, N. Costa, G. Ramalho, D. Gašević, 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), ACM, 2025, pp. 93–103. doi:10.1145/3706468.3706481.

[3] P. Šindelář, O. Bojar, S. Ettejjari, L. F. Vargas Madriz, M. Piacentini, K. Thomas, Overview of the PISA sensemaking task at the ELOQUENT 2026 lab for evaluating generative language model quality, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.

[4] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.

[5] Qwen Team, Qwen3 technical report, 2025. doi:10.48550/arXiv.2505.09388, arXiv:2505.09388.

[6] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with PagedAttention, in: Proceedings of the 29th Symposium on Operating Systems Principles (SOSP), 2023, pp. 611–626. doi:10.1145/3600006.3613165.

[7] S. Frohn, T. Burleigh, J. Chen, Automated scoring of short answer questions with large language models: Impacts of model, item, and rubric design, in: Proceedings of the 26th International Conference on Artificial Intelligence in Education (AIED 2025), volume 15882 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 44–51. doi:10.1007/978-3-031-98465-5_6.

[8] C. Zhao, M. Silva, S. Poulsen, Language models are few-shot graders, in: Proceedings of the 26th International Conference on Artificial Intelligence in Education (AIED 2025), volume 15880 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 3–16. doi:10.1007/978-3-031-98459-4_1.

[9] J. Gwozdz, A. Both, WSE-Research at BEA 2026 shared task: Zero-shot LLM-based short answer grading with rubric decomposition, in: Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA), Association for Computational Linguistics, 2026. To appear.

[10] A. Horbach, J. Pehlke, R. Laarmann-Quante, Y. Ding, Crosslingual content scoring in five languages using machine-translation and multilingual transformer models, International Journal of Artificial Intelligence in Education 34 (2024) 1294–1320. doi:10.1007/s40593-023-00370-1.

[11] F. Zehner, H. J. Shin, E. Kerzabi, A. Horbach, S. Gombert, F. Goldhammer, T. Zesch, N. Andersen, Down the cascades of Omethi: Hierarchical automatic scoring in large-scale assessments, in: Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025), Association for Computational Linguistics, 2025, pp. 660–671. doi:10.18653/v1/2025.bea-1.47.

[12] P. Šindelář, O. Bojar, Overview of the sensemaking task at the ELOQUENT 2025 lab: LLMs as teachers, students and evaluators, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), 2025. ArXiv:2507.12143.

[13] J. Karlgren, M. Barrett, O. Bojar, M. I. Engels, D. Fabre, S. Ettejjari, L. Goeuriot, R. R. Gunti, J. Mothe, P. Mulhem, M. Piacentini, L. F. Vargas Madriz, D. Schwab, P. Šindelář, G. Stampoulidis, K. Thomas, M. Vartampetian, Overview of ELOQUENT 2026: Shared tasks for evaluating generative language model quality, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[14] P. Šindelář, D. Slivka, C. Bouma, F. Prášil, O. Bojar, Training data generation for context-dependent rubric-based short answer grading, 2026. doi:10.48550/arXiv.2603.28537, arXiv:2603.28537.

[15] Gemma Team, Google DeepMind, Gemma 3 technical report, 2025. doi:10.48550/arXiv.2503.19786, arXiv:2503.19786.

[system]
You are scoring a student's answer to a reading comprehension question. Score ONLY based on the rubric criteria provided. Respond with ONLY the score number, nothing else.

[user]
Context: *{reading_passage}*
Question: *{question}*

Scoring rubric:
– Full Credit (FC, score 2): *{fc_description}*
– Partial Credit (PC, score 1): *{pc_description}*
– No Credit (NC, score 0): *{nc_description}*

Few-shot examples (k=3, TF-IDF retrieved)
Example 1: Answer: *{ex_1_answer}* → Score: *{ex_1_label}*
Example 2: Answer: *{ex_2_answer}* → Score: *{ex_2_label}*
Example 3: Answer: *{ex_3_answer}* → Score: *{ex_3_label}*

Target
Student answer: *{student_answer}*

Score this answer as exactly one of: 0, 1, 2

Figure 2: Rubric track prompt template (3-shot). System message Item context + rubric Retrieved examples Target answer . Placeholders in *italics*.

A. Prompt Templates

Figure 2 shows the **Rubric** track prompt template. The **Simple** prompt follows the same structure but omits the rubric and few-shot examples, replacing them with a 0–4 scale description.