Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno 6 n. 3 (2026)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

From Black-Box Automated Scoring to Engaging Feedback: A Docimological-Pedagogical Pipeline for Assessment with Local LLMs

Marco Giganti [1,*] and Marco Lazzari [1]

[1] Università degli Studi di Bergamo; marco.giganti@unibg.it; marco.lazzari@unibg.it
* Correspondence: marco.giganti@unibg.it

Abstract: The use of artificial intelligence in assessment raises questions about how automated judgements relate to learning goals and how students can use the feedback they receive. This paper presents the first stage of a broader research project on the assessment of written responses in Italian provided by local large language models (LLMs). We conducted a purposive narrative review of ten technical publications, retrieved through ACL Anthology, arXiv and publishers' websites before 1 October 2026, utilizing literature on feedback and assessment practices. The synthesis shows that scoring performance depends on the task, the human reference and the evaluation setting, while evidence concerning feedback use remains limited. These findings inform a proposed five-layer pipeline in which explicit criteria and independently rated responses provide the basis for semantic analysis, local-model evaluations and teacher review. The pedagogical rationale is to support judgements whose grounds can be verified and feedback that learners can use in their revision. As the pipeline has not yet been tested, we outline a validation program addressing scoring agreement, the quality of reviewed decisions, teacher workload and students' interpretation and use of feedback.

Keywords: automated short-answer grading; formative feedback; hybrid human-machine assessment

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the context of educational assessment, when interpreting an open-ended response, it is necessary to consider what the task was intended to elicit. An omitted condition, an incorrect causal inference or an unconventional, yet defensible, formulation acquires meaning in relation to the learning goals and the criteria used to assess them. Those criteria also allow a teacher to explain an evaluation and help the student decide how to revise the work. When an automated system contributes a score or a comment, its place in assessment depends on whether that contribution can be examined on the same grounds.

The issue becomes particularly relevant when feedback is expected to support learning. Carless and Boud (2018) describe feedback literacy through learners' capacity to understand information and use it to improve their work. Their account directs attention to what follows an evaluation, as students interpret a comment and decide how to act. In our design, engagement concerns that use of feedback, including questioning a judgement or explaining a revision. A promptly provided

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

comment can create an opportunity for such work, whose educational value depends on how learners respond.

Teachers' reported practices suggest that AI is less frequently utilized for assessment than in the preparation of didactical materials. In their analysis of OCSE TALIS 2024 data, Barabanti et al. (in press) examined 3,874 Italian lower-secondary teachers and 87,428 teachers in an EU comparison sample, obtaining weighted AI-use estimates of 25.4% and 33.6%, respectively. Among Italian AI users, summarizing information (70.2%) and generating lesson plans or activities (68.4%) were more common than assessing student work (27.4%). The item combining feedback and parent communication reached 32.3% (formative feedback was not considered separately). As TALIS refers to AI in a broad manner, these figures situate the assessment question without establishing the quality of generative applications or the reasons for non-use.

This paper develops the first stage of a broader project investigating how local large language models (LLMs) might assist the work of assessing open-ended short responses. We bring automated short-answer grading (ASAG), essay scoring (AES) and feedback research into dialogue with docimological concerns, asking what evidence would justify a judgement and what would make its feedback useful. The initial focus is on short-form responses in Italian to semi-structured prompts, for which expected content can be specified while allowing various formulations.

2. Materials and Methods

The study takes the form of a purposive narrative review supporting the design of subsequent empirical research. Readings used in the first stage of the project's designing were supplemented by full texts supplied by the research team and targeted web retrieval, completed on 1 October 2026. Titles, authors, DOIs and identifiers were used to locate primary records in ACL Anthology, arXiv, Springer Nature Link and MDPI. Queries included "Algorithmic fairness in automatic short answer scoring" and combinations of "automatic short answer" with "item response theory" and "Artificial Intelligence" and "deep neural networks". Searches combining "feedback literacy" and "student", together with "SANRA", located contextual literature. Retrieval of the papers supported the design problem without constituting a systematic review or systematically searching Scopus, Web of Science or ERIC.

Selection prioritized publications describing methods or evaluations of written-response assessment, particularly where Italian language, transfer across prompts, rubric use, cost or fairness were addressed. Ten technical publications were read in full, including version 1 of the Aggarwal and Gaggioli preprints. The 2024 preprint of the AI Assessment Scale and literature on feedback supplied educational context, while Barabanti et al. (in press) contributed to the description of teachers' practices outside the technical corpus. Product documentation was used solely to establish deployment options. Unrelated applications and claims available only through secondary summaries were excluded.

For each technical source, we examined response format and language, scoring target, inputs, evaluation procedures, outcomes and limitations. The synthesis compared scoring grounds and evaluation conditions to establish what teachers would need to inspect in assisted assessment. SANRA, the scale for the quality assessment of narrative review articles described by Baethge et al. (2019), informed

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

reporting; no score, formal risk-of-bias assessment or independent duplicate screening was undertaken. The heterogeneity of tasks and metrics precluded pooling. The review informs design choices and validation questions, with no new data or prototype evaluation reported here.

3. Results

3.1. Scoring evidence and the interpretation of responses

Neural models can learn relationships between texts and human scores, as Taghipour and Ng (2016) showed with a long short-term memory (LSTM) network that outperformed a strong baseline on quadratic weighted kappa1 without manually engineered features. For assessment support, the question is how such predictions relate to the evidence required by the task when a score has to be explained to a student.

An Italian research on short answers helps specify the problem. Menini et al. (2019) used fastText representations and a support vector machine to classify statistics responses in a dataset of 1,069 student-reference pairs, including constructed negations and paraphrases. F12 reached .909 and remained .878 on 77 new examination pairs, although balanced accuracy fell to .635 in a set where correct answers predominated. The difference cautions against judging the classifier by its aggregate F1 alone, while the treatment of both negation and mathematical symbols shows why semantic resemblance needs interpretation against the criterion. Statements regarding causality and association may use the same wording even though they express different concepts; distinguishing between them helps ensure that the score assigned reflects the concept being evaluated.

Measurement models approach the relationship between response evidence and scoring from another direction. Shibata and Uto (2022) integrated multidimensional item response theory into analytic essay scoring to improve interpretability, while Uto and Uchida (2020) combined an LSTM representation with ability estimated from objective items in the same test. In the latter study, 511 Japanese university students answered three short-answer questions and 44 objective items. Five-fold cross-validation yielded mean Pearson correlations of .754 for the conventional model and .762 for the proposed model; a conventional model with an added dense layer also averaged .762, and the ability-based model performed less well on Question 3, whose scores were strongly skewed. Since the results depend on the architecture and the question, and correlation does not imply absolute agreement, these results put the contribution of capability-related information into perspective. Using other responses to judge an answer also requires construct-based justification and a bias inspection; our initial prototype will therefore exclude student-level ability estimates.

[1] Quadratic Weighted Kappa (QWK) is a statistical measure that evaluates the agreement between two ordinal classification scales (in our case as human raters or machine learning models) while correcting for agreement that happens purely by chance.

[2] The F1 score is a machine learning metric that evaluates a classification model's performance by calculating the harmonic mean of precision and recall. The formula is: F1 = 2 x [(Precision x Recall) / (Precision + Recall)].

3.2. The conditions under which performance can be interpreted

Bidirectional Encoder Representations from Transformers (BERT) yielded results in assay evaluation similar to those of classical methods, but at a higher cost in terms of their configurations, which provided a reason to stick with simple comparators (Mayfield & Black, 2020). Misgna et al. (2025) also distinguish scoring familiar prompts from transfer to new questions. Their survey highlights the difficulty of explaining which features of an essay contribute to the predicted score. Research on feedback generation is also comparatively limited, which the authors suggest may partly reflect the scarcity of datasets containing feedback and annotations on specific assessment dimensions alongside essays and scores. For our project, these findings indicate the need to examine separately whether scoring performance holds when the prompt changes and whether the feedback helps students understand and revise their work.

Emirtekin's (2025) review of 49 studies published between 2018 and 2024 extends this picture. Human-model alignment varied, with inconsistency, hallucinations and weak explainability reported across both BERT encoders and generative models. The findings therefore need to be read in relation to task and model type, particularly given the predominance of English datasets and limited evidence on motivation and perceived fairness. In an Italian setting, Gaggioli et al. (2025) found weak agreement and limited stability when five models assessed 67 university essays against four criteria over three prompt replications. Individual human-rater records had not been retained, preventing estimation of inter-rater reliability. Without those records, model error becomes harder to distinguish from uncertainty in the human reference, which our design must preserve.

3.3. Feedback as an object of evaluation

The reference used to evaluate feedback requires similar care. EngSAF, developed by Aggarwal et al. (2024), contains 5,814 engineering responses with synthetic feedback generated using grading labels. The evaluation distinguishes between new responses to questions represented in the training data and responses to entirely new questions, allowing generalization across questions to be examined. In the limitations section of the first version, the authors also caution that synthetically generated feedback may contain bias, factual errors or overly generic content. Although this material enables work on feedback generation, label-informed comments cannot provide an independent human reference for another system. The question is whether a comment identifies a defensible difficulty and provides guidance usable in revision.

The judgement on which feedback rests also needs examination for fairness. In 38,722 German PISA responses, Andersen et al. (2025) found a small significant home-language disparity in their most accurate combination, a support vector machine with Robustly Optimized BERT Pretraining Approach[3] (RoBERTa) embeddings. Students primarily speaking another language at home (e.g. their parents' native language) were over-credited, with variation between items. Although

[3] Instead of changing the underlying network architecture, RoBERTa achieves superior Natural Language Processing (NLP) performance purely by optimizing how the model is trained.

this finding applies specifically to this model, it supports the idea that, in addition to accuracy, one should also examine who is being misclassified and in what way.

3.4. A proposed pipeline for the assessment of written responses in Italian

We propose five layers, summarized in Table 1, through which teachers initially review every score and comment. The sequence starts from the assessment intention because later outputs depend on what counts as evidence of learning. Criteria and rated examples make that intention available to the model while preserving the grounds on which its judgement can be examined.

Table 1. Proposed pipeline layers, teacher actions and validation priorities.

Layer	System contribution	Teacher action	Validation priority
Assessment design	Store task and score rules	Define construct and criteria	Learning-outcome alignment
Dataset and calibration	Organize separate ratings	Rate examples and resolve ambiguity	Agreement and coverage
Semantic diagnostics	Retrieve examples and flag differences	Inspect concepts and alternatives	Error detection and transfer
Local language model	Draft criterion-linked evaluations	Check against the response	Stability and factuality
Human review	Show evidence and record decisions	Approve, revise or reject	Decision quality and workload

Note. Calibration concerns scoring rules and observed errors rather than unsupported model confidence.

3.5. Establishing criteria and a human reference

The first layer correlates the score with the level of judgement required by a task. Structured tasks combine bounded prompts and closed responses, often allowing scoring keys, whereas semi-structured tasks elicit open responses to constrained prompts. Criterion grids can specify their required concepts and partial credit without equating understanding with a single formulation. Authentic answers would help calibrate acceptable alternatives, documented before evaluation, with the fluency that affects scores only when it falls within the construct. Unstructured tasks, leaving prompt and response open, would instead require analytic rubrics that describe the level of achievement based on defensible arguments.

The second layer establishes a human reference through responses stored with their prompts and independent ratings. At least two assessors would rate the validation subset before adjudication, preserving the uncertainty that precedes the assignment of an agreed-upon score. Partially correct answers would help clarify score boundaries. Training and evaluation sets must then remain separate, keeping paraphrases and synthetic variants with their source partition and excluding evaluation answers from prompts and retrieval indexes. This separation allows performance to be interpreted on unfamiliar material. Synthetically generated or modified responses could be used to examine how the system handles specific errors, such as a missing negation or an omitted condition. Estimates of scoring performance would instead be based on authentic student responses, comparing the system's evaluations with independent human ratings.

3.6. Relating semantic analysis to a proposed evaluation

The third layer would retrieve calibrated examples using sentence representations supplied by multilingual encoders. Similarity rules and supervised classifiers would serve as baselines, with models and preprocessing fixed within each comparison. Retrieval and clustering would help teachers inspect comparable formulations, whose resemblance must be interpreted against the criterion before inferring shared misconceptions or equal marks. Diagnostics would address consequential negations, numbers, units, causal claims, omitted conditions and contradictions. An entailment model could add a signal if separately evaluated in Italian, while keeping textual discrepancies distinct from their pedagogical interpretation.

The fourth layer would use the question, answer, criteria and examples to propose scores with relevant passages and a feasible next step, allowing abstention. Gemma 4 is a candidate because its models are downloadable for local applications (Google, n.d.), with configuration selected through task evaluation alongside memory and response-time constraints. Recording checkpoint, quantization and runtime would identify the configuration being compared. Numerical values and score ranges allow for direct verification, while explanations require a more in-depth analysis, since it cannot be assumed that a plausible description accurately reflects the model's reasoning.

Local deployment would allow for control over the processing location, provided that permissions, retention policies, and transmission controls also apply to logs, backups, and exports. The instructions contained in the text intended for students would remain assessment material and could not take precedence over the scoring instructions. Fine-tuning would require a sufficient amount of human-evaluated data and identifiable errors to justify the use of more complex approaches over simpler ones.

3.7. Teacher judgement and student use of feedback

The final layer returns the evaluation with its criterion, response passages and examples so the teacher can approve, revise or reject it and record the decision. Conflicting scores, unsupported claims and unusual answers would warrant closer examination, in which audit findings are evaluated in light of both detected and undetected errors, instead of relying exclusively on claims concerning the model's reliability. Routine cases would also be examined because fluent explanations may

make errors less noticeable under pressure due to time constraints. Initial judgements and corrections would reveal whether assistance improves decisions or induces mistakes otherwise avoided.

Following this review, students will receive feedback that connects their assessment to the evidence within their work and provides a clear path toward achievable revisions. Students would be told how AI contributed and could request reconsideration, extending to teachers' assessment practices the principle of explicit expectations discussed by Perkins et al. (2024) for student GenAI use. We would examine the formative value of this proposed extension through students' explanations of what they changed or why they rejected a suggestion. Those accounts would help would facilitate a distinction between a critical engagement with the judgement and an uncritical acceptance of the comment.

4. Discussion

4.1. Empirical validation of the assisted assessment process

The next stage would test whether assistance improves assessment decisions and provides useful feedback to students. Comparing an embedding baseline, a rubric-prompted local LLM and the pipeline would determine the specific contribution of retrieval and diagnostics. New answers to established prompts would be differentiated from novel prompts and subsequent cohorts, as demonstrating success in a single environment does not resolve the question of generalization. We would conduct repeated runs with fixed configurations to assess stability, incorporating fresh held-out checks following any modifications to the model or the rubric.

From a docimological perspective, disagreement needs interpretation in relation to the reference judgement and its consequences. Exact agreement and macro-F1 would describe categorical decisions, while mean absolute error would address score differences. Weighted kappa would be used for ordered categories where the weights can be justified. Human-human and human-model agreement with uncertainty intervals at item or criterion level would accompany analysis of over-crediting, under-crediting and partial-answer errors that correlation can conceal. In instances where ethically justified data and sufficient samples were accessible, subgroup reviews would assess the direction of errors at comparable levels of human scoring, accounting for factors such as language background. Small groups would limit interpretation, and attributes used for a review should not be included as model inputs.

The feedback would be reviewed independently to verify its accuracy and compliance with the criteria, and to provide guidance for the revision process, while the teachers' final decisions would be analyzed in light of the errors identified and those that were missed, as well as the percentage of cases flagged. Students' explanations and revisions would provide evidence of understanding, with transfer tasks helping distinguish conceptual improvement from copying. Carless and Boud (2018) give a basis for attending to this use of feedback, although the proposed measures require piloting. A counterbalanced comparison of teacher-only and assisted workflows on corresponding responses would track the time required for preparation, processing, checking and correcting, as well as errors and usability, while taking into account the work that is not captured by the inference time alone.

4.2. Educational implications and limits of the proposal

The pipeline may have value for a teaching team even in instances where the LLM provides minimal: to score prediction. Examining borderline responses against explicit criteria can reveal ambiguities in a task or acceptable alternatives not included in the original grid, making calibration part of assessment design. It remains to be determined whether automated assistance can support that discussion while maintaining a sustainable workload. For students, the corresponding opportunity consists of receiving feedback through an activity that allows them to interpret and act on it. Faster feedback can leave time for revision, but the learning opportunity depends on the teacher's guidance and the possibility of questioning the judgement. It is therefore necessary to examine engagement within the educational context where feedback is being utilized.

The selective review and heterogeneous publication types limit conclusions about preferred architectures, and preprint findings may change. No gain in learning, reliability or workload has yet been demonstrated. Extension to open argumentative writing, other disciplines or populations would require reconsidering the evidence for judgement. The project's eventual impact will be determined by the findings of future investigations conducted under the designated assessment conditions.

5. Conclusions

This first stage of our project identifies the necessary conditions to integrate local-model outputs into an assessment framework that is accessible for examination by both teachers and students. Criteria and independent ratings establish the reference, with semantic analysis and explanations contributing to judgements subject to review. Beginning with bounded responses in Italian allows those relationships to be studied. The next stage must establish whether teachers reach defensible decisions and students can use the feedback in subsequent work.

References

Aggarwal, D., Bhattacharyya, P., & Raman, B. (2024). *"I understand why I got this grade": Automatic short answer grading with feedback* (Version 1) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.12818

Andersen, N., Mang, J., Goldhammer, F., & Zehner, F. (2025). Algorithmic fairness in automatic short answer scoring. *International Journal of Artificial Intelligence in Education, 35*, 3128–3165. https://doi.org/10.1007/s40593-025-00495-5

Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA—a scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review, 4*, Article 5. https://doi.org/10.1186/s41073-019-0064-8

Barabanti, P., Giganti, M., & Viganò, R. (in press). Teachers' engagement with generative AI: Evidence from TALIS 2024. In *Proceedings of TEEM 2026*. Springer.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Emirtekin, E. (2025). Large language model-powered automated assessment: A systematic review. *Applied Sciences, 15*(10), Article 5683. https://doi.org/10.3390/app15105683

Gaggioli, A., Casaburi, G., Ercolani, L., Collovà, F., Torre, P., & Davide, F. (2025). *Assessing the reliability and validity of large language models for automated assessment of student essays in higher education* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.02442

Google. (n.d.). *Gemma 4 model overview*. Google AI for Developers. Retrieved October 1, 2026, from https://ai.google.dev/gemma/docs/core

Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring? In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, & T. Zesch (Eds.), *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 151–162). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.15

Menini, S., Tonelli, S., De Gasperis, G., & Vittorini, P. (2019). Automated short answer grading: A simple solution for a difficult task. In R. Bernardi, R. Navigli, & G. Semeraro (Eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)* (pp. 257–263). CEUR Workshop Proceedings. https://aclanthology.org/2019.clicit-1.39/

Misgna, H., On, B.-W., Lee, I., & Choi, G. S. (2025). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review, 58*, Article 36. https://doi.org/10.1007/s10462-024-11017-5

Perkins, M., Roe, J., & Furze, L. (2024). *The AI Assessment Scale revisited: A framework for educational assessment* (Version 1) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.09029

Shibata, T., & Uto, M. (2022). Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2917–2926). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.257/

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1193

Uto, M., & Uchida, Y. (2020). Automated short-answer grading using deep neural networks and item response theory. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education* (pp. 334–339). Springer. https://doi.org/10.1007/978-3-030-52240-7_61