Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD37946

An Evidence-Grounded Framework for Evaluating AI-Generated Educational Feedback

Xinqi Yang

Department of Science and Environmental Studies, The Education University of Hong Kong, Hong Kong, China
s1164617@s.eduhk.hk

Abstract. Educational feedback generated by artificial intelligence is becoming more and more popular in teaching, but there are still differences in how to evaluate its quality in existing research. This study examined 68 empirical studies and compared the criteria, methods, evaluators, and evidence used in these evaluations. The results show that the relevant evaluation can be roughly summarised into three dimensions: Evidence Validity, Rubric and Task Alignment, and Pedagogical Usefulness. Evidence Validity focuses on whether the feedback is accurate and reliable, and is based on the learner's actual work. Rubric and Task Alignment focuses on whether the feedback meets the task requirements, evaluation criteria, and learner responses. Pedagogical Usefulness focuses on whether the feedback is clear, specific, actionable, and can support revision and further learning. The reviewed studies also used a variety of evidence to judge the quality of feedback, including human judgement, learner perceptions, revision behaviour, learning outcomes, and automated metrics. Based on these findings, this study proposes an evaluation framework that links different dimensions of feedback quality with corresponding evidence to provide clearer evaluation ideas for subsequent research and educational practice.

Keywords: AI-generated feedback, educational feedback, feedback evaluation, evidence-grounded framework

1. Introduction

1.1. AI-generated feedback in educational contexts

Generative artificial intelligence (AI) is increasingly used for educational feedback. Recent studies have examined the application of feedback generated by ChatGPT and large language models (LLMs) in writing and other learning tasks [1-3]. One of the reasons is that students often need timely feedback, while teachers have limited time to provide individualised guidance. Research in this field also shows that people are increasingly paying attention to how AI feedback works, quality evaluation methods, learners' responses, and its possible impact on learning [3, 4].

Educational feedback research shows that effective feedback is not just about providing information to learners. Hattie and Timperley [5] linked feedback to learners' performance and understanding, and pointed out that the effect of feedback will vary depending on the type of

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD37946

feedback and how it is provided. Sadler [6] described formative feedback as a process in which learners understand the expected standard, compare it with their current performance, and take action to reduce the gap. Shute [7] also emphasised feedback that is targeted, clear, focused, and supportive of learning. Together, these studies suggest that feedback should be grounded in learner performance, aligned with relevant goals or criteria, and useful for improvement. These ideas also apply to AI-generated feedback.

1.2. Fragmentation in current evaluation practices

Existing research uses varied approaches to evaluate AI-generated feedback. Comparisons with teacher, peer, or expert feedback are common in authentic learning tasks [1, 8], while other work focuses on characteristics such as reliability and pedagogical quality [9, 10]. The evidence used for these judgements also varies. Rubric- and assessor-based approaches have been used to examine the reliability and rubric-related performance of GPT-generated feedback [9]. Reference-based analysis, by contrast, has been used to examine whether LLM-generated feedback remains faithful to student work [11]. Studies may therefore address similar questions about feedback quality using different evaluation approaches.

Current studies use different terms and focus on different parts of feedback quality. They also rely on different kinds of evidence, which makes comparison across studies less straightforward.

Each type of evidence answers a different question. Learner responses can show whether feedback is seen as useful, but they give limited information about its accuracy or fit with assessment criteria. Expert judgement and automated checks are more useful for examining whether the feedback is accurate and aligned with assessment criteria, although they say less about how learners understand or use it. Since these forms of evidence are usually analysed separately, it is necessary to synthesise them to illustrate the relationship between the main evaluation dimensions and the evaluation methods.

1.3. Research questions and contributions

This article adopts a structured comparative synthesis to examine how existing research evaluates the educational feedback generated by AI.

This research addresses the following three research questions:

RQ1: What dimensions have been used to evaluate the quality of AI-generated educational feedback?

RQ2: What methods, evaluators, and evidence sources have been used in existing evaluation practices?

RQ3: How can the identified dimensions and approaches be synthesised into an integrated framework for evaluating AI-generated educational feedback?

This study identifies the main quality dimensions and compares the methods and evidence sources used in the existing evaluation. On this basis, the study proposes a three-layer framework, including Evidence Validity, Rubric and Task Alignment, and Pedagogical Usefulness. The framework provides a clearer basis for evaluating the educational feedback generated by AI in different educational contexts.

DOI: 10.54254/2755-2721/2026.LD37946

2. Review methodology

2.1. Search scope and source selection

The literature retrieval starts with two recent reviews on AI-generated educational feedback: Zhang et al. [12] and Maier et al. [4]. The primary studies included in the two reviews formed the initial literature library. On this basis, more recent studies were identified through forward citation searching, and 108 candidate records were obtained. After removing duplicates and records for which full texts could not be obtained, 80 studies remained for screening, of which 68 met the inclusion criteria. This review adopts a review-seeded literature identification strategy and conducts a comparative synthesis on this basis.

2.2. Inclusion and exclusion criteria

The included studies need to examine the educational feedback generated by generative AI or LLMs and provide empirical evidence related to its evaluation. These forms of evidence include feedback quality, evaluation methods, evaluators, sources of evidence, learner responses, or learning outcomes directly related to feedback. The following studies are not included in the analysis: research that only focuses on the overall attitude towards AI; research that cannot distinguish feedback from broader AI-supported learning activities; and research that only involves scoring or ratings but does not provide instructional feedback. Studies using only traditional non-generative feedback systems and non-empirical studies will also be excluded. Review articles are not included in the final analysis sample.

2.3. Coding and comparative synthesis

These 68 studies are coded according to educational context, AI system, feedback tasks, quality criteria, evaluation methods, evaluators, sources of evidence and major findings. First, the original quality terms in each study were retained and compared. Then, in combination with the feedback concepts discussed in the introduction, the repeated or similar concepts are classified. Through this process, three broader dimensions are formed: Evidence Validity, Rubric and Task Alignment, and Pedagogical Usefulness. For RQ2, the evaluation method, the evaluator and the source of evidence are coded separately to distinguish what was evaluated from how it was assessed. The same study can be given multiple codes. Overlapping categories were checked again against the full texts. Finally, according to the main patterns and gaps found in the coding, the framework for answering RQ3 is further constructed. The complete list of the 68 included studies and their coding results is available upon request from the corresponding author.

3. Comparative analysis of feedback quality dimensions

Different studies do not define the quality of feedback exactly the same, and many studies involved multiple aspects at the same time. Evidence Validity appeared in 47 studies, Rubric and Task Alignment appeared in 19 studies, and Pedagogical Usefulness appeared in 58 studies. Among them, 46 studies covered more than one dimension.

DOI: 10.54254/2755-2721/2026.LD37946

3.1. Evidence validity

Evidence Validity mainly focuses on whether the feedback generated by AI can accurately interpret learner work, and whether the judgement is supported by existing evidence. Relevant studies often use terms such as accuracy, correctness, faithfulness, reliability and error detection performance, and also examine whether AI feedback can correctly identify problems and whether it produces false information or unsupported judgements.

This dimension is usually evaluated through human and reference-based comparisons. For example, Wang et al. [13] compared ChatGPT-generated feedback on student argumentation with human annotations and used precision and recall to evaluate its accuracy. Dai et al. [9] combined human judgements and rubric aspects to evaluate the feedback generated by GPT in terms of readability, effective feedback elements and reliability. Jia et al. [11] focused more directly on faithfulness, examining whether LLM-generated feedback remained grounded in student assignments rather than introducing unsupported content. Programming and mathematics studies used similar comparisons with expected answers, teacher judgements, or established system feedback.

Evidence Validity captures only part of overall feedback quality. Feedback may be correct and still remain too general, poorly matched to the task, or difficult for learners to use. Alignment and pedagogical usefulness may therefore also be relevant to evaluation.

3.2. Rubric and task alignment

Rubric and Task Alignment concerns how well feedback responds to the learner's work, task requirements, assessment criteria, or expected answer. It appeared in 19 studies. Relevant criteria included rubric alignment, criterion coverage, contextual appropriateness, relevance to student work, and consistency with expected feedback. Alignment was assessed through these criteria as well as comparisons with expected or reference feedback. These approaches distinguish task-specific feedback from feedback that may be correct but generic or poorly aligned.

Only a relatively small number of studies treated alignment explicitly. In other cases, similar concerns appeared under broader terms such as relevance, appropriateness, or usefulness. Alignment, however, remains conceptually distinct from validity. Correct information can still miss the relevant rubric criterion or fail to address the learner's difficulty. Treating alignment separately helps preserve this distinction in the evaluation framework.

3.3. Pedagogical usefulness

Pedagogical Usefulness appeared in 58 studies, making it the most common of the three dimensions. It concerns whether feedback helps learners understand their performance, decide what to do next, revise their work, or improve subsequent learning. Relevant research mainly describes this dimension in terms of usefulness, specificity, personalisation, clarity, actionability, supportive tone and explanatory value.

Studies examine this dimension from different angles. One aspect is the feedback itself, mainly whether it can provide clear and effective guidance for learners. Studies also examine how learners respond to this feedback, especially whether they actually used it in the revision process. Some studies have also further examined subsequent learning outcomes.

Positive evaluation of feedback does not necessarily mean that learners will actually use the feedback, nor does it necessarily mean that it will lead to better later performance. Some studies

DOI: 10.54254/2755-2721/2026.LD37946

reported favourable evaluations alongside limited revision use, while others found differences between learner perceptions and observable outcomes. Evidence about how learners understand and act on feedback can therefore add information beyond feedback ratings alone. Learner perceptions, uptake, and outcomes can provide evidence for pedagogical usefulness.

4. Evaluation methods and evidence sources

Studies also differed in how they evaluated feedback quality. Human-based evaluation appeared in 62 studies, learner- and outcome-based evidence in 66, and automated evaluation in 38. Automated evaluation was combined with at least one other evidence source in 37 studies, so these categories were not mutually exclusive. Surveys and ratings were most common, followed by automated metrics, revision analysis, interviews or focus groups, and qualitative coding.

4.1. Human-based evaluation

Human-based evaluation involved teachers, trained raters, peers, researchers, and experts. Methods included rubric-based ratings, comparative judgements, qualitative coding, and comparisons with teacher, expert, or peer feedback.

Human evaluation was common for dimensions requiring contextual or pedagogical judgement. Trained evaluators compared AI and human feedback on accuracy, clarity, personalisation, and actionability, while teachers or experts assessed alignment with expected answers, task requirements, or disciplinary standards. Banihashem et al. [1], for example, compared peer-generated and ChatGPT-generated feedback in essay writing.

Automated scores may provide limited evidence about contextual relevance, instructional appropriateness, and guidance quality. Human evaluation can address these aspects, although the resulting judgements are shaped by the rubric, evaluator expertise, and rating procedure. Using multiple, trained, or blinded raters may improve consistency. Human evaluation alone provides limited evidence about whether learners understand, use, or benefit from the feedback.

4.2. Learner- and outcome-based evaluation

Learner- and outcome-based evidence appeared in 66 studies and covered learner perceptions, behaviour or uptake, and revision or learning outcomes. Methods included surveys, ratings, revision analysis, interviews, pre/post tests, and interaction or log data. Learner perceptions appeared in 60 studies, behaviour or uptake in 55, and revision, performance, or learning outcomes in 60.

Learner perceptions captured whether AI feedback was viewed as useful, trustworthy, clear, supportive, or preferable to other sources. These responses describe experienced quality but provide limited evidence about objective accuracy or effectiveness. Behavioural evidence can therefore add information about how learners actually use the feedback.

Behavioural evidence included uptake, rejection, revision strategies, interaction traces, screen recordings, and processing time. These measures may help show whether learners act on feedback they view positively. Chen et al. [14] reported that second language (L2) learners rejected 45.9% of the ChatGPT-generated feedback. Rejection was often linked to mismatched expectations and the effort required to act on the feedback. Yu and Xie [15] found higher overall uptake for teacher feedback, although ChatGPT feedback was often comprehensive. These studies show that feedback quality, learner uptake, and revision provide related but distinct evidence.

DOI: 10.54254/2755-2721/2026.LD37946

Outcome-based evidence included revision quality, task performance, post-test scores, academic achievement, motivation, and self-regulated learning. Experimental and quasi-experimental studies examined these outcomes by comparing AI-generated feedback with teacher feedback, peer feedback, or no-feedback conditions. Outcome measures provide limited information about why particular feedback was effective or ineffective. Combining them with analysis of the feedback itself and learners' engagement can therefore provide a stronger basis for judging pedagogical usefulness.

4.3. Automated and hybrid evaluation

Automated evaluation appeared in 38 studies. In 37 of these studies, it was combined with another source of evidence. Researchers usually compare the feedback generated by AI with human annotations, expected answers or reference materials to test whether the feedback can correctly identify errors and whether its content is consistent with existing evidence.

It is used for research in the fields of mathematics, programming and argumentation, mainly to test the ability to identify errors and whether the feedback is consistent with the expected feedback. However, the information that automated assessment can provide is still limited, especially in showing how learners understand and use feedback in real learning situations.

In addition to automated evaluation, studies also combine human judgement and learner data to understand the quality of feedback more comprehensively. Automated analysis is mainly used to evaluate the accuracy and alignment of feedback, and learner data can reflect how learners understand and use feedback, and whether the feedback is related to learning progress.

5. Integrated evaluation framework

Existing studies focus on different aspects of feedback quality, and there is also a variety of evidence for evaluation. This framework integrates these research results into three related layers: Evidence Validity, Rubric and Task Alignment, and Pedagogical Usefulness. Each layer links the specific assessment focus with the corresponding evidence.

5.1. Framework structure and layer relationships

Evidence Validity mainly focuses on whether the feedback is accurate and whether it is based on learner work. Rubric and Task Alignment focuses on whether the feedback meets the task requirements and assessment criteria, and also looks at whether it really responds to learner work. Pedagogical Usefulness focuses more on learners, mainly examining whether feedback can support their understanding and improvement. Effective feedback should let learners know what needs to be improved and what action can be taken next. Learners' responses to feedback and later use of it can provide evidence for this layer of evaluation.

These three layers reflect the different aspects of feedback quality, but they are closely related. Accuracy can only explain a part of it, and task alignment and usefulness are also important aspects of evaluating the quality of feedback. Considering the three layers together can provide a more comprehensive judgement of feedback quality.

5.2. Operational components and evaluation questions

The operational structure of each layer is summarised in Table 1, which shows the evaluation methods and types of evidence that can be adopted. These questions are used to guide the assessment and can be adjusted according to different contexts.

DOI: 10.54254/2755-2721/2026.LD37946

Table 1. The specific structure of the proposed evaluation framework

Layer	Operational component	Evaluation question	Possible evidence
Evidence Validity	Accuracy / correctness	Does the feedback correctly interpret the learner's work?	Expert or teacher judgement; reference answer; precision/recall; annotated student work
Evidence Validity	Faithfulness / grounding	Are the claims supported by the learner's actual response or available evidence?	Human verification; hallucination analysis; source tracing; automated faithfulness metrics
Rubric and Task Alignment	Criterion alignment	Does the feedback address relevant rubric or assessment criteria?	Rubric mapping; teacher/expert rating; criterion coverage
Rubric and Task Alignment	Task relevance	Is the feedback specific to the task and learner response?	Student work comparison; expected feedback; contextual judgement
Pedagogical Usefulness	Clarity and specificity	Can learners understand what is wrong or needs improvement?	Learner ratings; expert rubric; qualitative analysis
Pedagogical Usefulness	Actionability	Does the feedback provide usable guidance for revision or next steps?	Revision analysis; uptake data; learner interviews
Pedagogical Usefulness	Support for improvement	Does the feedback support meaningful revision or subsequent learning?	Revision quality; post-tests; delayed outcomes; task performance

A method may be used for multiple layers of evaluation at the same time. Teacher ratings can help judge whether the feedback is correct and whether it is relevant to the task. Learner data can reflect how learners understand and use feedback, while automated assessment is more suitable for checking accuracy and alignment.

5.3. Coverage, applications, and boundaries

With the continuous development of AI feedback systems, the types of evidence required for evaluation may also change. In the early stage, the evaluation usually focuses more on whether the feedback is accurate and whether it matches the specific task. These aspects can be tested by student work, expected answers, rubrics or expert judgement. After entering the actual teaching environment, the focus of the evaluation will shift more to how learners respond to feedback and whether they will use the feedback to revise or improve their work. This helps avoid using the early technical assessment results directly as evidence of classroom effectiveness.

DOI: 10.54254/2755-2721/2026.LD37946

Using multiple sources of evidence can also find problems that may be ignored by a single method. For example, the feedback may get a high quality evaluation, but it may still have little effect on revision; or there are still problems with the accuracy or task alignment of the feedback, but the learners still evaluate it positively.

Not every study needs to include all three layers of evaluation. The specific scope of assessment can be determined according to the purpose of the assessment, the educational context and the stage of system development. The main function of the framework is to make clear what aspect of feedback quality is being evaluated and what evidence supports that judgement. It is mainly used to evaluate the educational feedback generated by artificial intelligence, and is not suitable for all applications of generative AI in education.

6. Conclusion

This review examines 68 empirical studies on the evaluation of AI-generated educational feedback. The study found that three main dimensions repeatedly appear in the existing literature: Evidence Validity, Rubric and Task Alignment, and Pedagogical Usefulness. Evidence Validity mainly focuses on whether the feedback is accurate and whether it is based on learner work. Rubric and Task Alignment focuses on whether the feedback meets the task requirements and assessment criteria. Pedagogical Usefulness focuses on whether learners can understand the feedback and use it to improve their work.

Not all of the reviewed studies use the same evidence. Some studies focus on the quality of feedback itself, while others examine how learners respond to feedback and the outcomes after the feedback was used. Different evaluation methods may draw different conclusions. Even if the feedback is not completely accurate or task alignment is limited, learners may still give positive responses. At the same time, the actual impact of some feedback that has received high quality evaluations on revision or subsequent performance is still limited. These differences show that the source of evidence is very important when evaluating the quality of feedback.

The framework proposed in this review integrates these research results and links each quality dimension with the evidence used to examine it. The framework is designed to help researchers make clearer assessment choices in different settings. It does not require the same assessment procedure for each study. More importantly, the purpose of the assessment should be clear, and the evidence used should match the feedback aspects examined.

Future research can further explore the interaction of these three dimensions in practical applications. It can also examine whether the combination of different assessment methods can provide stronger evidence in different educational contexts. The long-term effect also deserves further attention, especially how learners understand the feedback generated by AI and how they use this feedback over time.

References

[1] Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

[2] Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., ... & Oh, A. (2024, November). LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)* (pp. 284-293). https://aclanthology.org/2024.customnlp4u-1.21/

DOI: 10.54254/2755-2721/2026.LD37946

[3] Ba, S., Yang, L., Yan, Z., Looi, C. K., & Gašević, D. (2025). Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: A systematic literature review. *Computers and Education Open, 9*, 100284. https://doi.org/10.1016/j.caeo.2025.100284

[4] Maier, U., Seibold, M., & Klotz, C. (2026). LLM-generated formative feedback in education: A qualitative systematic literature review. *Computers and Education Open, 10*, 100374. https://doi.org/10.1016/j.caeo.2026.100374

[5] Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

[6] Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144. https://doi.org/10.1007/BF00117714

[7] Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102/0034654307313795

[8] Weidlich, J., Gotsch, F., Schudel, K., Marusic-Würscher, C., Mazzarella, J., Bolten, H., ... & Merki, K. M. (2025). Teacher, peer, or AI? Comparing effects of feedback sources in higher education. *Computers and Education Open, 9*, 100300. https://doi.org/10.1016/j.caeo.2025.100300

[9] Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., ... & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence, 7*, 100299. https://doi.org/10.1016/j.caeai.2024.100299

[10] Nazaretsky, T., Gabbay, H., & Käser, T. (2026). Can students judge like experts? A large-scale study on the pedagogical quality of AI and human personalized formative feedback. *Computers and Education: Artificial Intelligence, 10*, 100533. https://doi.org/10.1016/j.caeai.2025.100533

[11] Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., & Gehringer, E. (2024). On assessing the faithfulness of LLM-generated feedback on student assignments. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 491-499). https://doi.org/10.5281/zenodo.12729868

[12] Zhang, K., Li, S., & Huang, Y. (2026). Generative AI-driven feedback in higher education: A scoping review. *Assessment & Evaluation in Higher Education*, Advance online publication. https://doi.org/10.1080/02602938.2026.2617157

[13] Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity, 51*, 101440. https://doi.org/10.1016/j.tsc.2023.101440

[14] Chen, Z., Wei, W., Zhu, X., & Yao, Y. (2026). Unpacking the rejection of L2 students toward ChatGPT-generated feedback: An explanatory research. *ECNU Review of Education, 9*(1), 1-20. https://doi.org/10.1177/20965311241305140

[15] Yu, H., & Xie, Q. (2025). Generative AI vs. teachers: Feedback quality, feedback uptake, and revision. *Language Teaching Research Quarterly, 47*, 113-137. https://doi.org/10.32038/ltrq.2025.47.07