US-China Education Review A, May 2026, Vol. 16, No. 5, 275-282
doi: 10.17265/2161-623X/2026.05.001

AI-Empowered Teaching Quality Evaluation in Higher Education: A Three-Path Practical Framework

JIANG Jingjing, HU Shurui
Shanghai Jiao Tong University, Shanghai, China

This paper explores how artificial intelligence can be embedded into classroom teaching quality evaluation in higher education and proposes a practical framework organized around three layers: perception, analysis, and service. In response to the limited coverage of expert observation, fragmented evaluative evidence, and the weak translation of feedback into improvement, three implementation paths are developed: AI-based video analysis, AI-supported data analysis, and AI-enabled intelligent assistants. Video analysis generates traceable process evidence from classroom activities, teacher-student behaviors, and interaction patterns. Data analysis integrates AI-generated indicators, expert observation, student feedback, and teacher self-evaluation to support comparison, diagnosis, and trend tracking. Intelligent assistants provide policy interpretation, result explanation, task reminders, and improvement support for teachers, supervisors, and administrators. The paper further proposes a five-stage implementation process from data preparation to follow-up review. It argues that AI should not replace human evaluators in making final judgments, but should provide explainable and reviewable evidence for a teaching quality evaluation system moving from one-time scoring toward continuous improvement.

Keywords: artificial intelligence, teaching quality evaluation, higher education, video analysis, data analysis, intelligent assistant

Introduction

Background and Problem Statement

Classroom teaching lies at the core of talent cultivation in higher education. At present, however, classroom teaching quality evaluation practices widely adopted in Chinese universities still face several limitations. First, expert observation coverage is restricted because universities offer a large number of heterogeneous courses while trained experts are limited. Second, evaluative evidence remains fragmented: expert comments, student evaluation of teaching questionnaires, and teacher self-evaluations are often collected at different times and interpreted separately, while student ratings may also be affected by bias. Third, feedback often lacks timeliness, continuity, and actionable diagnostic support; expert scores may cluster at the upper end or be influenced by individual preferences, and teachers may not receive support when instructional problems first emerge.

In recent years, the rapid advancement of artificial intelligence has opened new possibilities for evaluating classroom teaching quality in higher education. Computer vision can support the recognition of classroom behaviors, postures, and interaction patterns; natural language processing enables the semantic parsing and

JIANG Jingjing, M.Ed., Engineer, Educational Technology Center, Shanghai Jiao Tong University, Shanghai, China.
HU Shurui, M.Ed., Engineer, Educational Technology Center, Shanghai Jiao Tong University, Shanghai, China.

sentiment analysis of classroom discourse and student feedback; large language models, with their strong capacities for understanding and generation, support the intelligent analysis of instructional content and the automatic drafting of evaluation reports; and learning analytics integrates multimodal data to build process-oriented evaluation models. The convergence of these technologies is driving classroom evaluation away from traditional, outcome-based, and experience-driven judgments toward a paradigm that is data-driven, process-oriented, and increasingly intelligent.

Purpose and Significance

This study is positioned as a practical framework study rather than an empirical test of a single AI tool. Organized along the three layers of "Perception–Analysis–Service", it constructs a pathway framework based on AI-based video analysis, AI-supported data analysis, and AI-enabled intelligent assistants. The focus is to clarify how different AI capabilities can be connected with existing university quality assurance processes. The framework may help universities improve teaching supervision efficiency, generate traceable evidence, reduce purely impression-based evaluation, and establish a closed-loop process of "evidence collection-feedback-improvement-follow-up".

Literature Review

Evolution of Teaching Quality Evaluation in Higher Education

The evaluation of teaching quality in higher education has gradually shifted from single-source judgment to evidence-based and multi-source evaluation. In earlier institutional practice, student evaluation of teaching (SET) was one of the most widely used instruments for assessing instructional effectiveness. Marsh (1987) provided influential evidence that SET has a multidimensional structure and acceptable reliability at the aggregate level, which helped establish its legitimacy in university quality assurance systems. However, as SET results became increasingly linked to promotion, accountability, and performance management, concerns about their validity and potential biases became more prominent. Spooren, Brockx, and Mortelmans (2013) reviewed the state of research on SET and argued that, despite its extensive use, unresolved issues remain regarding construct validity, teacher-related bias, and the interpretation of results for formative improvement. Uttl, White, and Gonzalez (2017) further challenged the assumption that SET ratings are valid proxies for teaching effectiveness, showing through meta-analysis that student ratings are not consistently associated with student learning outcomes.

This critique has encouraged a broader movement toward developmental and evidence-based approaches to teaching evaluation. Berk (2005) proposed a triangulated model that integrates student ratings, peer review, self-evaluation, teaching portfolios, learning outcomes, and other sources of evidence, emphasizing that no single measure can adequately capture the complexity of teaching. Similarly, Pianta and Hamre (2009) argued that standardized classroom observation, represented by instruments such as the Classroom Assessment Scoring System (CLASS), can transform classroom interactions into systematic and feedback-oriented evidence. The Measures of Effective Teaching project further demonstrated that combining classroom observations, student surveys, and achievement gains yields more stable and informative indicators than any single source alone (Kane & Staiger, 2012). Therefore, the contemporary development of teaching quality evaluation is not a simple replacement of SET, but a transition toward multi-actor, multi-evidence, and improvement-oriented systems. The main practical constraint is that traditional human observation remains costly, infrequent, and difficult to scale, which creates a clear rationale for the introduction of AI-supported evaluation.

AI Technologies Applied to Classroom Observation and Feedback

Early applications of artificial intelligence in educational evaluation were primarily associated with learner profiling, predictive analytics, automated assessment, adaptive learning systems, and intelligent tutoring. In a systematic review of AI applications in higher education, Zawacki-Richter, Mar ń, Bond, and Gouverneur (2019) identified assessment and evaluation as a major application area, while also noting that much of the field remains technology-driven and insufficiently connected to pedagogical theory, educator perspectives, and ethical reflection. As classroom data have expanded from questionnaires and achievement scores to audio, video, discourse transcripts, and behavioral logs, AI-supported evaluation has begun to move from outcome scoring toward process-oriented understanding. For instance, Demszky et al. (2021) developed a computational framework for measuring conversational uptake in teacher-student interactions, showing that discourse features such as acknowledgement, revoicing, and building on student contributions can be operationalized and associated with instructional quality.

Recent advances in multimodal learning analytics and large language models have further expanded the possibilities of automated classroom observation and feedback. Ramakrishnan et al. (2023) developed ACORN, a multimodal machine learning system that uses audio, visual, and facial-expression features to estimate CLASS dimensions such as positive climate and negative climate; its predictive performance was reported to approach the level of agreement among human coders. This provides empirical support for the feasibility of AI-assisted classroom observation. At the same time, human-AI complementarity remains a central design principle. Holstein, McLaren, and Aleven (2019) argued that AI-enhanced classroom systems should support teachers' orchestration and decision-making rather than replace professional judgment. In the context of generative AI, Wang and Demszky (2023) examined whether ChatGPT could score classroom instruction and generate actionable feedback. Their findings suggest that large language models can produce rubric-relevant responses, but their suggestions often lack novelty and context-sensitive insight. Taken together, existing studies indicate that AI can support classroom behavior recognition, discourse analysis, multimodal observation, and automated feedback, but its outputs still require expert interpretation, calibration, and ethical oversight. This is especially important in higher education, where classroom formats, disciplinary standards, privacy requirements, and teacher agency vary substantially across contexts.

Analytical Framework and Application Scenario

The Perception-Analysis-Service Framework

The framework proposed in this paper is based on the logic that classroom teaching quality evaluation should not stop at the production of scores. A complete AI-empowered evaluation system needs to cover evidence generation, diagnostic interpretation, and improvement support. Accordingly, the framework is divided into three layers. The perception layer uses AI-based video and audio analysis to transform classroom behaviors, interaction events, and teaching activities into measurable process evidence. The analysis layer integrates AI-generated indicators with expert observation, student feedback, and teacher self-evaluation to form a more comprehensive picture of teaching quality. The service layer delivers evaluation standards, analytical findings, reminders, and improvement suggestions to teachers, supervisors, students, and administrators through intelligent assistants and role-based interfaces.

Practical Application Scenario in Universities

In a typical university context, the framework can be embedded into the existing teaching quality assurance system rather than replacing it. Classroom videos generated by smart classrooms can be analyzed after class; expert supervisors can review AI-flagged sessions; student evaluation data can be incorporated at mid-semester or at the end of the semester; teachers can use diagnostic reports to adjust instruction; and administrators can monitor course-level and department-level trends. Expert supervision remains responsible for professional judgment, while AI provides broader coverage, more timely evidence, and more accessible feedback.

Constructing Practical Pathways for AI-Empowered Classroom Teaching Quality Evaluation in Higher Education

Figure 1 summarizes the relationship between the three paths, their functional positions, data sources and outputs.

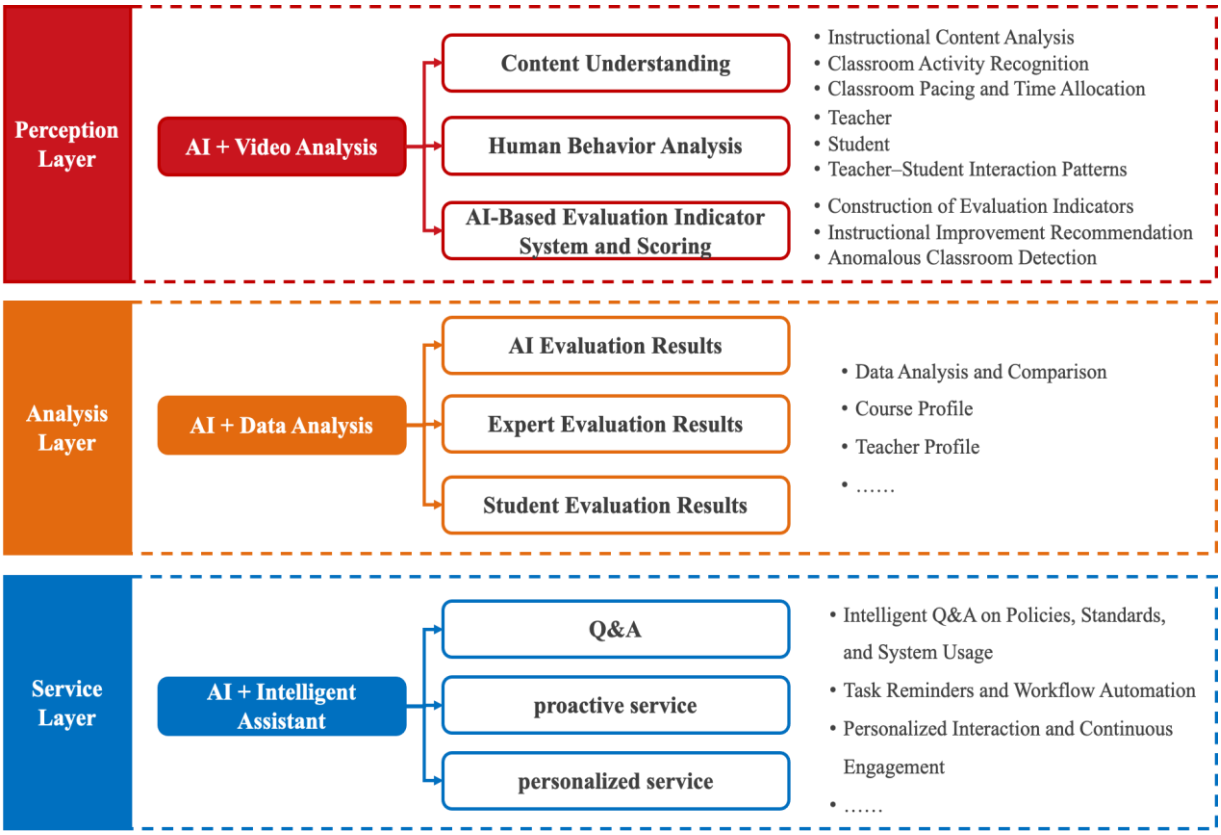

Figure 1. The three paths.

Path 1: AI-Based Video Analysis

Positioned at the perception layer, this path employs AI as a classroom observer that extracts structured evidence from classroom video. Multimodal models can identify dominant instructional events such as lecturing, questioning, discussion, practice, demonstration, or student presentations, and can generate an activity map for each lesson. By combining automatic speech recognition with visual cues, the system can also analyze lesson pacing, time allocation, and the way instructional content is represented through slides, board work, or demonstrations.

Computer vision and audio analysis can further describe teacher movement, gesture, eye contact, verbal exchange, student participation, and interaction coverage. These features are then mapped to a hierarchical indicator system covering content delivery, interaction quality, student engagement, and classroom organization. The output is a structured report containing preliminary dimensional scores, key moments, and retrievable video clips. Since automated inference has clear limitations, these scores should be used as traceable evidence for expert review rather than as final judgments.

Path 2: AI-Supported Data Analysis

At the analysis layer, AI functions as a data analyst that integrates heterogeneous evaluation evidence. The framework consolidates AI-generated video scores, expert observation ratings and comments, student evaluation of teaching questionnaires, open-ended feedback, and, where available, teacher self-evaluations. Because these sources differ in scale and timing, a mapping scheme is needed to project them onto shared evaluation dimensions and meaningful aggregation windows.

Once aligned, the data enable several forms of comparison. AI and expert evaluations can be compared to identify convergence and divergence; AI observations and student perceptions can be used together to reveal hidden issues, such as frequent but superficial interaction; longitudinal analysis can track the development of individual teachers; and cross-sectional analysis can build course or department profiles. Analytical outputs should be differentiated by audience: teacher reports emphasize diagnosis and improvement suggestions, departmental reports support course group management, and institutional reports focus on trends, warning signals, and resource allocation.

Path 3: AI-Enabled Intelligent Assistants

At the service layer, AI operates as an intelligent service agent for teachers, supervisors, administrators, and students. Built on large language models grounded in institutional knowledge bases, the assistant can answer questions about evaluation policies, indicator definitions, report interpretation, appeal procedures, and system operation. For example, a teacher may ask why a particular interaction score is low, and the assistant can connect the explanation with relevant video segments and comparative data.

Beyond question-answering, the assistant can provide task reminders and workflow support, such as prompting experts to complete observation reports, students to submit SET questionnaires, and teachers to review feedback. It can also deliver personalized improvement resources based on diagnostic results and collect user feedback on both the evaluation results and the system itself. Through such interaction, evaluation becomes less like a periodic administrative task and more like a continuous support mechanism for teaching improvement.

Implementation Process of the AI-Empowered Evaluation Framework

To make the above three paths operational, universities may adopt a gradual five-stage process.

The first stage is data preparation and standard setting. Universities need to review existing rubrics, expert observation forms, SET questionnaires, and teaching quality management documents, and extract common dimensions such as teaching content, classroom organization, student engagement, interaction quality, and teaching effectiveness. At the same time, rules for classroom video collection, storage, access, and use should be established before AI analysis begins.

The second stage is intelligent classroom perception. Classroom video and audio data are processed through speech recognition, activity recognition, behavior detection, and multimodal analysis. The aim is not to make a final judgment, but to transform the teaching process into structured evidence such as activity timelines, key interaction moments, and preliminary indicator scores.

The third stage is multi-source data integration and diagnosis. AI-generated indicators are compared with expert comments, SET results, and teacher self-evaluation. When different sources converge, conclusions become more credible; when they diverge, the difference itself becomes a topic for further review.

The fourth stage is feedback delivery and improvement support. Teachers receive individual diagnostic reports, supervisors receive key video segments for review, departments receive course group profiles, and administrators receive institutional trends and warning signals.

The fifth stage is follow-up review. After teachers adjust instructional design or interaction strategies, subsequent classroom data can be used to observe whether improvement has occurred.

In this way, AI-empowered evaluation becomes a cycle of "evidence-diagnosis-feedback-action-review" rather than a one-time scoring activity.

Governance, Ethics, and Risk Control

The implementation of AI-empowered classroom teaching evaluation must be accompanied by clear governance mechanisms. Because classroom video and behavioral data involve teachers and students directly, the system should follow the principles of informed notification, transparent use, minimum necessity, limited access, and accountable review. Universities should clarify the purpose, scope, retention period, and user permissions of data use, and restrict access to original video clips to authorized personnel.

Teacher agency is also essential. If AI evaluation is perceived as surveillance or as an automatic ranking tool, teachers may resist the system and its formative value will be weakened. Therefore, AI-generated scores should be treated as diagnostic evidence rather than final conclusions. Teachers should have the right to view the evidence behind scores, provide explanations, request human review, and appeal inappropriate interpretations. In addition, the indicator system, scoring logic, and limitations of AI models should be made understandable to users, and AI outputs should be regularly compared with expert judgment to identify possible bias across disciplines and classroom types.

Limitations and Future Development

This paper proposes a practical framework and implementation process, but it does not claim that AI-based classroom evaluation can be directly applied in all universities without adjustment. Its feasibility depends on classroom recording systems, data platforms, evaluation rubrics, and technical support teams. Universities with limited digital infrastructure may need to begin with small-scale pilots before expanding to school-wide implementation.

AI analysis of classroom behavior also has inherent limitations. Observable behaviors such as attention-related visual cues, speaking turns, and interaction frequency cannot fully represent learning quality, teaching depth, or students' cognitive engagement. Different disciplines and classroom types may require different indicators; the same interaction pattern may have different meanings in a large lecture, a clinical training session, or a graduate seminar. Future research should develop discipline-sensitive indicators, examine the alignment

between AI-generated evidence and expert judgment, and evaluate whether AI-supported feedback leads to sustained improvement in teaching practice.

Conclusion

This paper has examined three persistent limitations in classroom teaching quality evaluation in higher education: insufficient observation coverage, fragmented evaluative evidence, and weak translation of feedback into improvement. In response, a three-layer framework of perception, analysis, and service, and developed three practical paths: AI-based video analysis, AI-supported data analysis, and AI-enabled intelligent assistants, is proposed.

The contribution of this framework lies in connecting AI capabilities with the actual workflow of university teaching quality assurance. Video analysis provides traceable process evidence, data analysis integrates multi-source information for diagnosis, and intelligent assistants translate results into accessible services. Through the five-stage process of data preparation, intelligent perception, multi-source diagnosis, feedback delivery, and follow-up review, teaching evaluation can gradually shift from one-time scoring toward a continuous improvement loop.

The central argument of this paper is that AI should be positioned as an augmenting mechanism rather than an autonomous evaluator. AI-generated scores should not replace human judgment, but should provide explainable, traceable, and reviewable evidence. Only when technical implementation is combined with governance arrangements, teacher agency, algorithmic transparency, and institutional trust can AI-empowered evaluation become a sustainable pathway for improving classroom teaching quality in higher education.

References

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48-62.

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 1638-1653). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.130

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education: 20th International Conference, AIED 2019* (pp. 157-171). Springer. https://doi.org/10.1007/978-3-030-23204-7_14

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains (MET Project Research Paper). Bill & Melinda Gates Foundation.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253-388. https://doi.org/10.1016/0883-0355(87)90001-2

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119. https://doi.org/10.3102/0013189X09332374

Ramakrishnan, A., Zylich, B., Ottmar, E., LoCasale-Crouch, J., & Whitehill, J. (2023). Toward automated classroom observation: Multimodal machine learning to estimate CLASS positive climate and negative climate. *IEEE Transactions on Affective Computing, 14*(1), 664-679. https://doi.org/10.1109/TAFFC.2021.3059209

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598-642. https://doi.org/10.3102/0034654313496870

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wang, R. E., & Demszky, D. (2023). Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 626-667). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.53

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0