Beyond Task-Based Evaluation: Toward Situated Benchmarks for Educational LLMs

Anonymous ACL submission

Abstract

This paper argues that evaluation of educational LLMs should move beyond decontextualized task-based benchmarks toward situated benchmarks centered on pedagogical action. Prior work in human-computer interaction and the learning sciences has long emphasized situated use, user context, and interaction processes; our focus is narrower and complementary: how these concerns should reshape benchmark design for educational LLMs. Existing evaluations already cover question generation, knowledge answering, error diagnosis, feedback generation, and teaching analysis, but their dominant unit remains the single input-output episode, making it difficult to compare whether a model genuinely supports learning progress, instructional judgment, or subsequent revision in context. Educational activity is processual, context-dependent, and development-oriented, so explanations, feedback, and hints can only be evaluated properly in relation to learner state, interaction history, and pedagogical goals. We review a landscape still dominated by task-based paradigms and argue, from educational conceptions, educational values, and research methodology, that next-generation educational benchmarks should make pedagogical function in context a core target. We further propose layered, scalable, and benchmarkable situated protocols built through simulation-first development and follow-up empirical calibration, extending evaluation beyond task quality to contextual fit and downstream consequences for learning and teaching decisions.

1 Introduction

Large language models (LLMs) are rapidly expanding across educational settings, and benchmarks have expanded with them. Recent work already covers question generation, knowledge understanding, error diagnosis, feedback generation, and teaching analysis (Chen et al., 2024; Zhou et al., 2025a; Hou et al., 2024; Rittle-Johnson et al., 2025; Chu et al., 2026; Wang et al., 2025). In this area, benchmarks do more than passively measure performance. They influence how the field understands educational capability, sets optimization targets, and judges progress.

Despite this expanding thematic scope, the dominant paradigm remains task-based (Chen et al., 2024; Hou et al., 2024; Su et al., 2025; Chu et al., 2026). Whether the object of evaluation is a test item, a piece of instructional feedback, or a classroom analysis, most benchmarks still center on a single output and ask whether it is correct, clear, complete, or aligned with preset requirements (Fu et al., 2024; Raihan et al., 2024; Rittle-Johnson et al., 2025; Wang et al., 2025). This paradigm has clear advantages: explicit criteria, low implementation cost, reproducibility, and easy cross-model comparison. But it mainly measures local output quality and says much less about support for learning progress, teacher judgment, or later instructional decisions.

We argue that the next generation of educational LLM benchmarks should evaluate not only whether a model can complete an isolated task, but also whether it can take pedagogically meaningful action in concrete or simulated teaching contexts.

This claim builds on longstanding concerns in human-computer interaction and the learning sciences with situated use, context, and interaction processes (Suchman, 1987; Dourish, 2004; Brown et al., 1989; Lave and Wenger, 1991). Our question is how to translate those concerns into educational LLM benchmarks that remain reusable and comparable while representing learner state, interaction history, pedagogical goals, and the consequences of instructional action.

This translation is especially important in education. Education is organized around learning goals, and learners enter each activity with prior understanding, error histories, and interaction his-

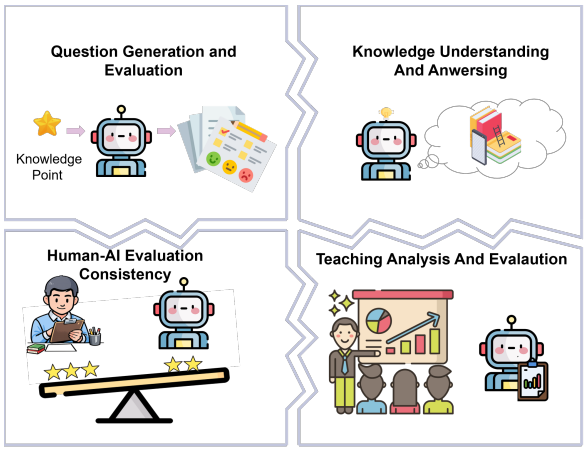

Figure 1: Task families are expanding, but contextualized evaluation remains limited.

tories. The pedagogical value of an explanation, a piece of feedback, or a classroom analysis therefore depends not only on its content, but also on whether it fits the teaching situation and helps advance later learning or teaching adjustment (Brown et al., 1989; Lave and Wenger, 1991; Macina et al., 2023; Mitton et al., 2026; Chu et al., 2026).

For this reason, task-based evaluation alone cannot capture the goals educational LLMs are meant to serve (Jiang et al., 2025; Rittle-Johnson et al., 2025; Chu et al., 2026; Weitekamp et al., 2025; Li et al., 2026). Well-formed feedback does not mean a student can revise understanding, and correct answers do not mean a system can scaffold a misconception appropriately. If benchmarks continue to reward correctness, completeness, and fluency above all else, optimization will stay focused on static, locally measurable indicators, while learner agency, error repair, knowledge transfer, and support for teacher judgment remain marginal. As LLMs move further into tutoring, assessment design, feedback, lesson support, and classroom analysis, this divergence becomes harder to ignore.

Task-based benchmarks remain core infrastructure and a practical starting point, but in education they are insufficient on their own. The evaluative focus should gradually shift from high-quality outputs to high-quality pedagogical action in context.

Against this background, this paper makes three contributions:

1. **Landscape synthesis.** We review the mainstream task-based paradigm in educational LLM benchmarking and synthesize existing evaluations across major educational task families.

2. **Theoretical argument.** We show, from the perspectives of educational conception, educational value, and research methodology, why educational benchmarks should go beyond decontextualized task evaluation.

3. **A practical roadmap.** We propose a layered situated benchmarking framework developed through simulation-first iteration and follow-up empirical calibration.

We do not offer a finished unified benchmark or argue for an immediate shift to long-horizon empirical evaluation. This is a benchmarking position: if educational LLM evaluation is to align with practice, tasks must be put back into educational context, and pedagogical action in context must become a core target.

2 The Current Paradigm: Task-Based Benchmarks and Integrated Educational Evaluations

Educational LLM benchmarking has developed a relatively clear evaluation ecology. Existing work can be grouped into five broad directions: question generation and assessment, knowledge understanding and application, human-LLM evaluation alignment, teaching analysis and evaluation, and integrated educational evaluation. Most studies still use a single item, a single feedback turn, a classroom fragment, or one interaction as the basic unit of evaluation, so comparison remains organized around correctness, consistency, alignment, or aggregate quality (Fu et al., 2024; Raihan et al., 2024; Su et al., 2025; Chu et al., 2026). A smaller set of recent studies has begun to place models inside multi-turn tutoring, simulated classrooms, or instructional workflows (Weitekamp et al., 2025; Jeon et al., 2026; Wei et al., 2025; Li et al., 2026). Integrated educational evaluation has therefore emerged as a horizontal extension across task families, although the corresponding environments and protocols remain unstable. Figure 2 summarizes this landscape.

2.1 Teaching-material preparation

Teaching-material preparation forms one of the earliest mature directions in educational benchmarking. The central concern is whether a model can generate high-quality educational materials and evaluate their instructional quality in a stable way. Work in this area spans multiple-choice

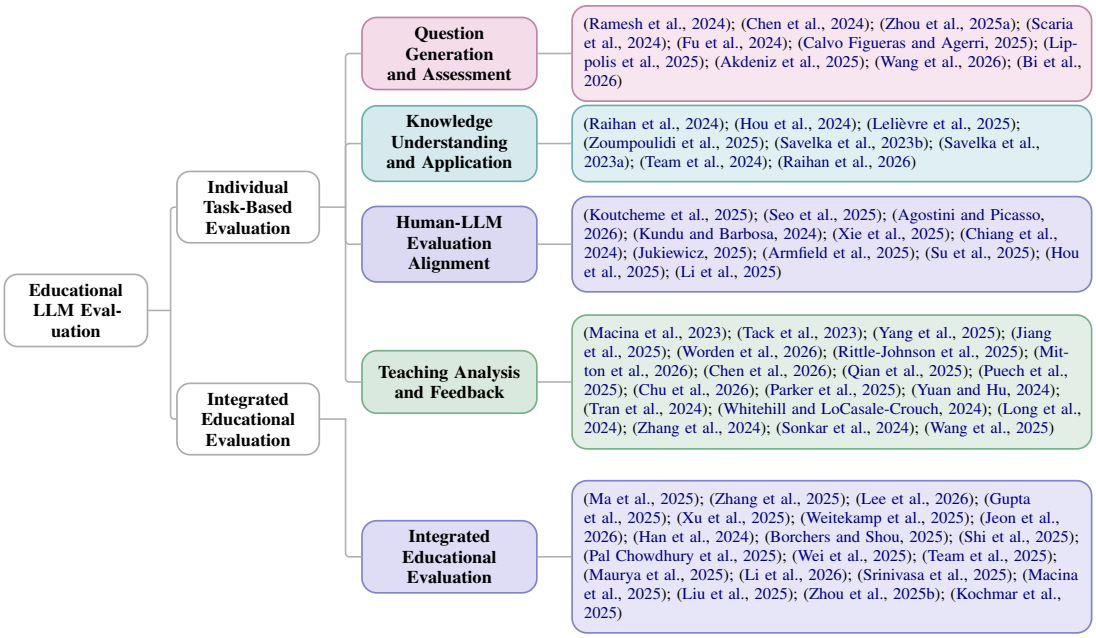

Figure 2: A taxonomy of educational LLM evaluation.

item quality, decomposed questioning capability, answer-to-question generation, multi-dimensional quality protocols, and later extensions built around Bloom levels, cognitive depth, critical question generation, and competency-oriented questioning (Ramesh et al., 2024; Chen et al., 2024; Zhou et al., 2025a; Fu et al., 2024; Calvo Figueras and Agerri, 2025; Lippolis et al., 2025; Akdeniz et al., 2025; Wang et al., 2026). Recent work has further extended educational content generation to the multimodal dimension, automating the creation of illustrative materials alongside text-based items (Bi et al., 2026), and to instructional-design agents that prepare broader teaching materials and activities (Jeon et al., 2026). Together, these studies show that teaching-material evaluation is moving from surface text quality toward alignment with cognitive levels, competency targets, multimodal content forms, and instructional-design requirements.

2.2 Knowledge understanding and application

Knowledge understanding and application form another major line of work. The question here is whether a model can complete curricular tasks, explain knowledge content, and maintain stable performance on subject-specific or course-level tasks. Existing benchmarks cover K–12 evaluation, programming education, pedagogical knowledge, learning-oriented explanation, and Bloom-aligned explanation quality (Raihan et al., 2024; Hou et al., 2024; Savelka et al., 2023b; Raihan et al., 2026; Lelièvre et al., 2025; Zoumpoulidi et al., 2025; Team et al., 2024). This line of work shows that educational evaluation is expanding from answer correctness to course-task performance, pedagogical knowledge, and explanation quality, while still largely treating the evaluated unit as an individual task or response.

2.3 Human–LLM alignment in educational assessment

Human-LLM evaluation alignment extends the model's role from answerer to evaluator. These studies ask whether model judgments can remain aligned with human evaluation. The relevant settings include programming-feedback judgment, educational feedback assessment, essay and assignment scoring, multi-granular writing evaluation, instructor calibration, and human-in-the-loop designs (Koutcheme et al., 2025; Seo et al., 2025; Agostini and Picasso, 2026; Kundu and Barbosa, 2024; Armfield et al., 2025; Su et al., 2025; Li et al., 2025). The core issue is whether models can produce judgments that teachers can accept, compare, and revise.

2.4 Instructional analysis and diagnosis

Instructional analysis and evaluation shift the evaluation target toward student thinking, instructional dialogue, and classroom materials. Studies in this line examine tutoring dialogue, multimodal tutoring, guided instruction, knowledge tracing, and

pedagogical grounding, showing that evaluation has expanded from answer correctness to teaching-related understanding (Macina et al., 2023; Yang et al., 2025; Jiang et al., 2025; Worden et al., 2026; Long et al., 2024). Misconception diagnosis is the most concentrated subfield, with evaluation extending from right-or-wrong answers to error types, reasoning steps, and thought structures (Rittle-Johnson et al., 2025; Mitton et al., 2026; Chen et al., 2026). Work on feedback quality, pedagogical steering, pedagogical alignment, and classroom-material analysis further widens the boundary of this task family (Puech et al., 2025; Chu et al., 2026; Tran et al., 2024; Sonkar et al., 2024; Wang et al., 2025).

2.5 Integrated educational evaluations

Integrated educational evaluation is the most notable recent horizontal extension in educational LLM benchmarking. Broad benchmark suites already integrate cognitive levels, subjects, competence structures, and languages (Ma et al., 2025; Zhang et al., 2025; Lee et al., 2026; Gupta et al., 2025), while newer work extends this logic to educational scenarios, tutor-student testbeds, multi-agent frameworks, syllabus-grounded teaching, indirect teaching-based evaluation, and shared-task protocols (Xu et al., 2025; Weitekamp et al., 2025; Wei et al., 2025; Li et al., 2026; Zhou et al., 2025b; Kochmar et al., 2025). These efforts show a clear shift toward integrated evaluation, but they remain methodologically heterogeneous and often rely on synthetic tasks, proxy students, LLM judges, or turn-level criteria. The field therefore still lacks a unified framework for assessing pedagogically meaningful action in context, motivating the layered situated benchmarks proposed in this paper.

3 Why Task-Based Evaluation Is Not Enough: Three Positions

3.1 Conceptual position: Education is processual, situated, and developmental

From a conceptual standpoint, the limits of task-based evaluation are intrinsic. Education is a continuous developmental process shaped by learning goals and learner state (Brown et al., 1989; Lave and Wenger, 1991). New learning activities are built on prior understanding, error histories, interaction histories, and affective conditions (Macina et al., 2023; Mitton et al., 2026). Educational action therefore has to be interpreted within a learning trajectory, and judgment should rest on how later understanding and later action change (Weitekamp et al., 2025; Li et al., 2026). Task-based evaluation treats educational capability as a set of local tasks that can be decomposed and evaluated independently. That treatment does not match the real organization of educational practice.

Educationally important actions depend on concrete context, and this is clear in question generation, knowledge support, and feedback. Whether a question is appropriate depends on whether it matches the learner's current level (Akdeniz et al., 2025), exposes a key misconception, and helps move learning forward (Wang et al., 2026). Whether an explanation works depends on whether it addresses the learner's immediate conceptual bottleneck, fits prior understanding, and helps the learner grasp underlying concepts and relations (Zoumpoulidi et al., 2025; Team et al., 2024). Whether feedback is effective depends on whether it helps the learner revise understanding, remain engaged, and continue participating in later learning (Macina et al., 2023; Chu et al., 2026). Once detached from concrete teaching situations, these tasks become difficult to judge appropriately.

Success on a single task does not imply instructional effectiveness. A model may generate a question, a feedback message, or an explanation, but those outputs alone do not establish pedagogical value. That judgment still depends on the organization of question sequences, the process by which student understanding is revised, and the fit between explanation and the learner's cognitive starting point.

3.2 Value position: Task optimization may drift from educational value

From a value-based standpoint, subtask evaluation is necessary and feasible, but its coverage is limited. Benchmarks do not only measure performance; they also shape research attention and optimization direction. If evaluation continues to revolve around correctness, completeness, fluency, and immediate helpfulness (Fu et al., 2024; Chu et al., 2026), optimization will keep concentrating on single-turn task completion, while higher-quality understanding, revision, transfer, and support for teacher judgment struggle to become core targets (Maurya et al., 2025; Li et al., 2026).

This drift is especially visible in education. Rapid access to standard answers may compress the space for independent thinking, for surfacing

misunderstanding, and for repairing error (Rittle-Johnson et al., 2025; Mitton et al., 2026). Stronger direct assistance may also increase learner dependence and further crowd out teacher-student and student-student interaction (Jiang et al., 2025; Liu et al., 2025). Likewise, more fine-grained classroom labels, summaries, or feedback texts are not automatically more educationally valuable (Whitehill and LoCasale-Crouch, 2024; Wang et al., 2025). What matters is whether they genuinely support sustained learning and subsequent instructional decision-making.

Local task metrics therefore cannot stand in for educational value on their own. Educational benchmarks also need to ask whether learning is advanced, whether understanding is repaired, whether transfer becomes possible, and whether interaction provides useful grounds for teacher judgment (Weitekamp et al., 2025; Li et al., 2026). Only by returning single tasks to concrete teaching situations can benchmarking identify educational values that matter but are difficult to read directly from static outputs.

3.3 Methodological position: Educational benchmarks should reconnect with empirical research

From a methodological standpoint, task-based benchmarks have clear strengths. They are highly standardized, comparatively inexpensive, and easy to compare across models, which makes them useful infrastructure for model comparison. But the methodological concerns of educational research go beyond whether a static output satisfies a formal criterion. Educational research is concerned with whether instructional designs, feedback strategies, and interaction structures have meaningful effects in real or near-real settings (Brown et al., 1989; Lave and Wenger, 1991). If benchmarking remains confined to static proxy indicators, ecological validity, teacher participation, process evidence, and downstream learning outcomes stay outside the evaluative frame (Weitekamp et al., 2025; Li et al., 2026).

A better path is to treat more context-sensitive benchmarking as an intermediate layer between scalable model comparison and empirical educational research. It preserves standardized comparison, rapid screening, and low-cost iteration while moving targets and evidence closer to issues central to educational research, including conceptual repair, knowledge transfer, learning engagement,

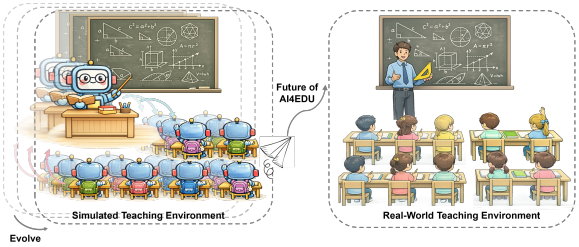

Figure 3: A bridge from controllable virtual environments to calibrated real-world validation.

teacher judgment, multi-turn coherence, and downstream learning outcomes (Macina et al., 2023; Chu et al., 2026; Weitekamp et al., 2025; Li et al., 2026). Otherwise, benchmarking will keep refining static proxy indicators.

4 Design Principles for Situated Benchmarks

Situated benchmarks therefore require coordinated redesign of the evaluation target, protocol, implementation path, and metric system. We organize the proposal around four principles: teaching action as the unit of evaluation, layered protocols, simulation-first development with follow-up empirical calibration, and metrics for task quality, contextual fit, and learning progression, as illustrated in Figure 4.

4.1 Evaluation unit: From isolated outputs to teaching actions in context

In a situated benchmark, the evaluation target is no longer a single output detached from context. It is a teaching action interpreted within a concrete instructional process. In educational settings, the quality of a judgment depends not only on the content itself, but also on the contextual conditions in which the action occurs, the person toward whom it is directed, and the influence it may have on subsequent learning (Dourish, 2004). At minimum, evaluation therefore needs to connect contextual conditions, teaching action, and subsequent evidence, although the precise content of each component should be defined by task rather than forced into a universal template. Benchmarking no longer evaluates only output quality. It also evaluates action timing, scaffold design, interaction coherence, and support for later learning.

4.2 Evaluation method: Layered protocols for instructional processes

Situated benchmarks are better served by layered protocols. Educational judgment is inherently multi-level: it concerns whether a task is completed, whether the action fits the current context, and whether it helps move later learning forward (Brown et al., 1989; Lave and Wenger, 1991). We therefore propose a three-layer structure.

- **Layer 1: Task Quality.** This layer evaluates whether the output is accurate, clear, coherent, and structurally sound.
- **Layer 2: Contextual Fit.** This layer evaluates whether the action is appropriate given the learner's level, interaction history, learning goals, and teacher-side constraints.
- **Layer 3: Learning Progression.** This layer evaluates whether the action supports later revision, transfer, self-explanation, or other forms of educational advancement.

This structure separates different kinds of educational judgment, so the benchmark no longer returns only an overall score. It can instead show whether the problem lies in task quality, contextual judgment, or learning progression, thereby providing more interpretable and more targeted evidence.

4.3 Implementation path: Simulation-first iteration and small-scale empirical calibration

Situated benchmarks also need an implementation path that can be sustained. Educational settings are highly complex, so benchmark design has to preserve educational relevance while remaining iterable, reusable, and shareable. We therefore propose a gradual strategy: simulation-first iteration followed by empirical calibration.

Virtual environments provide a low-cost prototyping space for benchmark construction. In multi-agent classroom simulations, structured tutoring traces, or teaching-design workflows, researchers can test whether an evaluation unit is reasonable, whether an interaction structure is stable, and whether a scoring protocol is operable. Virtual environments also make it easier to control variables, accumulate shared cases, and accelerate iteration.

Small-scale empirical calibration then returns tasks and protocols to real educational conditions.

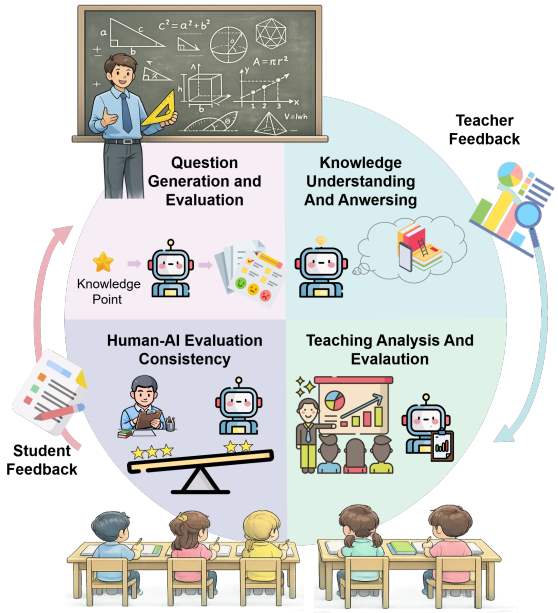

Figure 4: A process-oriented framework for situated educational evaluation.

Teacher review, limited student interaction, and reanalysis of authentic classroom fragments can test whether the design still holds and provide external validity support. Together, virtual environments and empirical calibration offer a workable balance among educational relevance, research feasibility, and construction cost.

4.4 Metric system: Task quality, contextual fit, and learning progression

The metric system should correspond to the three-layer structure above, but it should not be treated as a fixed scale for every task. Task quality, contextual fit, and learning progression provide organizing principles. Different tasks may adopt different metric forms, evidence sources, and implementation strategies, so long as they make model performance across the three layers legible and comparable.

5 From Task Families to Situated Extension

Situated benchmarking becomes operational only when its design principles enter established educational task families. The more practical path is not to build a wholly separate evaluation system, but to reconstruct evaluation units, layered judgments, and follow-up evidence on top of existing task families. Building on the taxonomy in Section 2, we organize these extensions into four directions: from item quality to learning pathways, from correct answering to history-sensitive understanding support,

from one-shot feedback to repair trajectories, and from static analysis to instructional decision support (Brown et al., 1989; Lave and Wenger, 1991).

5.1 From item quality to learning pathways

Current teaching-material benchmarks still tend to assess materials as isolated items or outputs, using criteria such as clarity, curricular alignment, difficulty, distractor quality, and instructional relevance (Fu et al., 2024; Akdeniz et al., 2025). Such criteria are necessary but incomplete, since the educational value of a material also depends on its position in a learning sequence and the learner activity it elicits.

The situated extension of teaching-material preparation should therefore move from judging isolated outputs to evaluating the organization of learning pathways. Possible targets include progressive item sequences, hint sequences, misconception-exposing prompts, differentiated questioning, and supporting materials. The emphasis shifts from surface material quality to the role of instructional materials in structuring learning progress.

5.2 From knowledge understanding to history-sensitive understanding support

Knowledge-understanding and answering benchmarks mainly evaluate whether a model can answer correctly, provide reasonable explanations, and maintain stable knowledge performance (Hou et al., 2024; Zoumpoulidi et al., 2025). These tasks reflect foundational disciplinary competence, but knowledge support in education is an ongoing process. What learners often need is not an abstract standard answer, but an explanation that connects with prior understanding, addresses the source of error, and supports later transfer.

The situated extension of this task family should therefore shift from one-shot correctness to history-sensitive understanding support. The key issue is no longer simply whether the model knows the answer, but whether it can use prior errors, earlier responses, learning trajectory, and current confusion to provide appropriate explanation, re-representation, and prompting (Brown et al., 1989; Lave and Wenger, 1991). Static accuracy alone is therefore insufficient.

5.3 From one-shot feedback to repair trajectories

Feedback generation and error diagnosis are already close to authentic instructional support, yet mainstream benchmarks still focus on whether a single feedback response is accurate, clear, helpful, and well matched to student error. They are therefore closer to instructional process than simple answering tasks, but still organized around one-turn feedback quality.

The key situated extension is to place feedback back within the revision process. A more educationally meaningful benchmark should ask whether the model can support a repair trajectory across multiple turns, including whether it chooses an appropriate scaffold, follows up productively after partial revision, avoids premature answer disclosure, and supports deeper error reconstruction. The evaluation unit thus shifts from a single response to a continuous trajectory of revision support.

5.4 From static teaching analysis to instructional decision support

Teaching analysis and evaluation have expanded benchmarking from student-side tasks to instructional understanding, but mainstream assessment still focuses mainly on whether the analytical result itself is correct. For educational use, however, analysis only becomes fully valuable when it enters the chain of instructional decision-making.

The situated extension of this task family should therefore move from static analysis to instructional decision support. The focus is not only whether a model can correctly describe a classroom episode, but whether it can help teachers identify ideas worth probing, prioritize key misconceptions, and propose adjustments fitted to current class conditions. Evaluation thus expands from analytical correctness to decision-support utility.

5.5 A realistic prioritization strategy

Not all of these extensions need to be advanced simultaneously. The most practical starting points are tasks already close to process-oriented instructional support and easier to connect to follow-up evidence in virtual environments, such as feedback generation, hint generation, misconception diagnosis, teacher-support analysis, and item-sequence design. Tasks that require longer time spans or higher ecological realism, such as full classroom management, long-term learning-path tracking, or complex course implementation, can follow later.

6 Discussion: Feasibility, Limitations, and Implementation Constraints

To make situated benchmarks buildable evaluation schemes, three issues need clarification: the status of task-based benchmarks, the implementation conditions of situated evaluation, and the limitations of the present claim (Suchman, 1987; Dourish, 2004).

6.1 Task-based benchmarks still matter

This paper does not reject task-based benchmarks. They make it possible to decompose complex educational capability into stable units, compare systems in standardized ways, and reproduce results at relatively low cost (Chen et al., 2024; Hou et al., 2024). Even after a situated turn, they still serve as infrastructure. If a system cannot maintain basic correctness in knowledge, feedback, or analysis, it is difficult to discuss contextual fit or later effects. Task-based results also provide the lower-level constraints that make situated comparison interpretable. Situated benchmarks therefore complement rather than replace task-based evaluation, and are better understood as a higher-layer evaluation built on task-based results.

6.2 Construction conditions and implementation pathway for situated evaluation

The main challenges of situated benchmarks are higher cost, slower construction, and greater subjectivity. Compared with single-item answering or one-turn feedback, instructional action in context is harder to standardize and depends more heavily on history, role relations, outcome format, and human judgment (Suchman, 1987; Dourish, 2004). It also faces practical constraints involving rating consistency, scenario sampling, and the collection of follow-up evidence.

Feasibility comes from staged construction rather than immediate reproduction of an entire classroom. Evaluation units and protocols can first be iterated in multi-agent simulations, structured traces, classroom scripts, and instructional-design workflows, then calibrated through teacher review, small-scale student participation, and authentic classroom fragments. This approach avoids hard-coding one teaching situation as the only valid standard. A buildable benchmark does not require full classroom realism at every stage, but explicit, revisable simplifications open to external checking. Virtual environments provide iterability, while empirical calibration keeps benchmarks from becoming closed systems detached from reality.

Situated benchmarks should therefore not begin by pursuing exhaustive scenario coverage or a once-and-for-all protocol. A more realistic strategy is to start with teaching episodes that have clearer structure, role boundaries, and follow-up evidence, then expand to more complex settings after local protocols stabilize. They should also report where judgment remains fragile, such as scorer disagreement, scenario sensitivity, or limited follow-up evidence, rather than compressing uncertainty into a single score.

6.3 Limitations of the present position

This position also has concrete limitations. We do not argue that all educational benchmarks should become long-horizon, multi-turn, heavily empirical evaluations; for many foundational capabilities, compact task-based evaluation remains more appropriate. Nor do we argue that only real-classroom evidence has value: virtual environments, structured traces, and near-authentic teaching settings can all ground situated benchmarks, although none replaces educational research itself (Cobb et al., 2003; Messick, 1995). More importantly, the simulation-first route may still misrepresent real classrooms when proxy learners, synthetic traces, or LLM judges diverge from authentic interaction. We also do not claim that situated benchmarking can immediately deliver a unified standard. Because disciplines, grade levels, and pedagogical goals define good instructional action differently, a more realistic outcome is the gradual formation of partially shared principles, locally comparable protocols, and expanding task families rather than a single global score or leaderboard.

7 Conclusion

This paper argues that educational LLM benchmarking should move beyond decontextualized task evaluation toward situated evaluation of teaching processes. Because education is processual, context-dependent, and developmental, single-output quality alone cannot determine whether a model supports learning or instruction. We therefore propose a situated turn that evaluates pedagogical action in context through layered evidence and simulation-first development with empirical calibration. The next step is not to abandon tasks, but to return them to educational context and stabilize standards that better reflect educational practice.

References

Daniele Agostini and Federica Picasso. 2026. Evaluation of large language models' educational feedback in higher education: Potential, limitations, and implications for educational practice. *arXiv preprint arXiv:2602.02519*.

Hasan Akdeniz, Tyler Clark, and Julia Link Roberts. 2025. Can ai generate questions aligned with bloom's taxonomy? a framework for gifted education to support teachers. *Journal of Advanced Academics*, 36(4):671–694.

Derek Armfield, Eason Chen, Asilbek Omonkulov, Xinyi Tang, Jionghao Lin, Erik Thiessen, and Kenneth Koedinger. 2025. Avalon: A human-in-the-loop llm grading system with instructor calibration and student self-assessment. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*, pages 111–118. Springer.

Shuzhen Bi, Mingzi Zhang, Zhuoxuan Li, Xiaolong Wang, Keqian Li, and Aimin Zhou. 2026. Eduillustrate: Towards scalable automated generation of multimodal educational content. *arXiv preprint arXiv:2604.05005*.

Conrad Borchers and Tianze Shou. 2025. Can large language models match tutoring system adaptivity? a benchmarking study. In *Artificial Intelligence in Education*, volume 15878 of *Lecture Notes in Artificial Intelligence*, pages 407–420. Springer.

John Seely Brown, Allan Collins, and Paul Duguid. 1989. Situated cognition and the culture of learning. *Educational Researcher*, 18(1):32–42.

Blanca Calvo Figueras and Rodrigo Agerri. 2025. Benchmarking critical questions generation: A challenging reasoning task for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5635–5652, Suzhou, China. Association for Computational Linguistics.

Xinghe Chen, Naiming Liu, and Shashank Sonkar. 2026. Malrulelib: Large-scale executable misconception reasoning with step traces for modeling student thinking in mathematics. *arXiv preprint arXiv:2601.03217*.

Yuyan Chen, Chenwei Wu, Songzhou Yan, Panjun Liu, Haoyu Zhou, and Yanghua Xiao. 2024. Dr. academy: A benchmark for evaluating questioning capability in education for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3138–3167.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Seongyeub Chu, Jongwoo Kim, and Munyong Yi. 2026. Feedeval: Pedagogically aligned evaluation of llm-generated essay feedback. *arXiv preprint arXiv:2601.04574*.

Paul Cobb, Jere Confrey, Andrea diSessa, Richard Lehrer, and Leona Schauble. 2003. Design experiments in educational research. *Educational Researcher*, 32(1):9–13.

Paul Dourish. 2004. What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1):19–30.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. Qgeval: Benchmarking multi-dimensional evaluation for question generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11783–11803, Miami, Florida, USA. Association for Computational Linguistics.

Vansh Gupta, Sankalan Pal Chowdhury, Vil'em Zouhar, Donya Rooein, and Mrinmaya Sachan. 2025. Are large language models for education reliable across languages? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 612–631, Vienna, Austria. Association for Computational Linguistics.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024. 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)*, pages 284–293. Association for Computational Linguistics.

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. 2024. E-eval: A comprehensive chinese k-12 education evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7753–7774, Bangkok, Thailand. Association for Computational Linguistics.

Zhaoyi Hou, Alejandro Ciuba, and Xiang Li. 2025. Improving llm-based automatic essay scoring with linguistic features. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 41–65. PMLR.

YoungHoon Jeon, Suwan Kim, Haein Son, Sookbun Lee, Yeil Jeong, and Unggi Lee. 2026. Isd-agentbench: A comprehensive benchmark for evaluating llm-based instructional design agents. *arXiv preprint arXiv:2602.10620*.

Mei Jiang, Houping Yue, Bingdong Li, Hao Hao, Ying Qian, Bo Jiang, and Aimin Zhou. 2025. Sid: Benchmarking guided instruction capabilities in stem ed-

ucation with a socratic interdisciplinary dialogues dataset. *arXiv preprint arXiv:2508.04563*.

Marcin Jukiewicz. 2025. A systematic comparison of large language models for automated assignment assessment in programming education: Exploring the importance of architecture and vendor. *arXiv preprint arXiv:2509.26483*.

Ekaterina Kochmar, Kaushal Kumar Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Ana"is Tack, and Justin Vasselli. 2025. Findings of the bea 2025 shared task on pedagogical ability assessment of ai-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Charles Koutcheme, Nicola Dainese, Arto Hellas, Sami Sarsa, Juho Leinonen, Syed Ashraf, and Paul Denny. 2025. Evaluating language models for generating and judging programming feedback. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*.

Anindita Kundu and Denilson Barbosa. 2024. Are large language models good essay graders? *arXiv preprint arXiv:2409.13120*.

Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge.

Unggi Lee, Sookbun Lee, Heungsoo Choi, Jinseo Lee, Haeun Park, Younghoon Jeon, Sungmin Cho, Minju Kang, Junbo Koh, Jiyeong Bae, Minwoo Nam, Juyeon Eun, Yeonji Jung, and Yeil Jeong. 2026. Openlearnlm benchmark: A unified framework for evaluating knowledge, skill, and attitude in educational large language models. *arXiv preprint arXiv:2601.13882*.

Maxime Lelièvre, Amy Waldock, Meng Liu, Natalia Vald'es Aspillaga, Alasdair Mackintosh, Mar'ia Jos'e Ogando Portela, Jared Lee, Paul Atherton, Robin A. A. Ince, and Oliver G. B. Garrod. 2025. Benchmarking the pedagogical knowledge of large language models. *arXiv preprint arXiv:2506.18710*.

Hang Li, Yucheng Chu, Kaiqi Yang, Yasemin Copur-Gencturk, and Jiliang Tang. 2025. Llm-based automated grading with human-in-the-loop. *arXiv preprint arXiv:2504.05239*.

Zheng Li, Siyao Song, Jingyuan Ma, Rui Li, Ying Zeng, Minghao Li, and Zhifang Sui. 2026. Teachbench: A syllabus-grounded framework for evaluating teaching ability in large language models. *arXiv preprint arXiv:2601.21375*.

Anna Sofia Lippolis, Minh Davide Ragagni, Paolo Ciancarini, Andrea Giovanni Nuzzolese, and Valentina Presutti. 2025. Bench4ke: Benchmarking automated competency question generation. *arXiv preprint arXiv:2505.24554*.

Ying Liu, Can Li, Ting Zhang, Mei Wang, Qiannan Zhu, Jian Li, and Hua Huang. 2025. Discerning minds or generic tutors? evaluating instructional guidance capabilities in socratic llms. *arXiv preprint arXiv:2508.06583*.

Yun Long, Haifeng Luo, and Yu Zhang. 2024. Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, 9(1):60.

Guoqing Ma, Jia Zhu, Hanghui Guo, Weijie Shi, Yue Cui, Jiawei Shen, Zilong Li, and Yidan Liang. 2025. Edueval: A hierarchical cognitive benchmark for evaluating large language models in chinese education. *arXiv preprint arXiv:2512.00290*.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221, Suzhou, China. Association for Computational Linguistics.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Joshua Mitton, Prarthana Bhattacharyya, Digory Smith, Thomas Christie, Ralph Abboud, and Simon Woodhead. 2026. Misconception diagnosis from student-tutor dialogue: Generate, retrieve, rerank. *arXiv preprint arXiv:2602.02414*.

Sankalan Pal Chowdhury, Terry Jingchen Zhang, Donya Rooein, Dirk Hovy, Tanja K"aser, and Mrinmaya Sachan. 2025. Educators' perceptions of large language models as tutors: Comparing human and ai tutors in a blind text-only setting. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 356–374, Vienna, Austria. Association for Computational Linguistics.

Michael J. Parker, Caitlin Anderson, Claire Stone, and YeaRim Oh. 2025. A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education*, 35:444–481.

Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26291–26311.

Keyang Qian, Yixin Cheng, Rui Guan, Wei Dai, Flora Jin, Kaixun Yang, Sadia Nawaz, Zachari Swiecki, Guanliang Chen, Lixiang Yan, and Dragan Ga sevi'c. 2025. Dean of llm tutors: Exploring comprehensive and automated evaluation of llm-generated educational feedback via llm feedback evaluators. *arXiv preprint arXiv:2508.05952*.

Nishat Raihan, Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, Christian Newman, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Cseprompts: A benchmark of introductory computer science prompts. In *Foundations of Intelligent Systems (ISMIS 2024)*, volume 14670 of *Lecture Notes in Computer Science*, pages 45–54. Springer.

Nishat Raihan, Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, Mohammed Latif Siddiq, Christian Newman, Tharindu Ranasinghe, Joanna C. S. Santos, and Marcos Zampieri. 2026. On the performance of large language models on introductory programming assignments. *Journal of Intelligent Information Systems*, 64:239–263.

Aninditha Ramesh, Arav Agarwal, and Jacob Doughty. 2024. A benchmark for testing the capabilities of llms in assessing the quality of multiple-choice questions in introductory programming education. In *Proceedings of the 2024 ACM Virtual Global Computing Education Conference (SIGCSE Virtual 2024)*.

Bethany Rittle-Johnson, Rebecca Adler, Kelley Durkin, L. Burleigh, Jules King, and Scott Crossley. 2025. Detecting math misconceptions: An ai benchmark dataset. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 20–24, Pittsburgh, Pennsylvania, USA. National Council on Measurement in Education (NCME).

Jaromir Savelka, Arav Agarwal, Marshall An, Chris Bogart, and Majd Sakr. 2023a. Thrilled by your progress! large language models (gpt-4) no longer struggle to pass assessments in higher education programming courses. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1 (ICER 2023)*, Chicago, IL, USA. ACM.

Jaromir Savelka, Arav Agarwal, Christopher Bogart, and Majd Sakr. 2023b. Large language models (gpt) struggle to answer multiple-choice questions about code. *arXiv preprint arXiv:2303.08033*. A conference version also appears in SCITEPRESS/ICALT-indexed proceedings; this entry cites the arXiv version pending venue preference.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In *Artificial Intelligence in Education. AIED 2024*, volume 14830 of *Lecture Notes in Computer Science*, pages 165–179. Springer.

Hyein Seo, Taewook Hwang, Jeesu Jung, Hyeonseok Kang, Hyuk Namgoong, Yohan Lee, and Sangkeun Jung. 2025. Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences*, 15(2):671.

Yao Shi, Rongkeng Liang, and Yong Xu. 2025. Educationq: Evaluating llms' teaching capabilities through multi-agent dialogue framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32799–32828, Vienna, Austria. Association for Computational Linguistics.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard G. Baraniuk. 2024. Pedagogical alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Rakshith S. Srinivasa, Zora Che, Chen Bo Calvin Zhang, Diego Mares, Ernesto Hernandez, Jayeon Park, Dean Lee, Guillermo Mangialardi, Charmaine Ng, Ed-Yeremai Hernandez Cardona, Anisha Gunjal, Yunzhong He, Bing Liu, and Chen Xing. 2025. Tutorbench: A benchmark to assess tutoring capabilities of large language models. *arXiv preprint arXiv:2510.02663*.

Jiamin Su, Yibo Yan, Fangteng Fu, Zhang Han, Jingheng Ye, Xiang Liu, Jiahao Huo, Huiyu Zhou, and Xuming Hu. 2025. Essayjudge: A multi-granular benchmark for assessing automated essay scoring capabilities of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6363–6389, Vienna, Austria. Association for Computational Linguistics.

Lucy A. Suchman. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, Cambridge.

Ana"is Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The bea 2023 shared task on generating ai teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Ankit Anand, Avishkar Bhoopchand, Brett Wiltshire, Daniel

Gillick, Daniel Kasenberg, Eleni Sgouritsa, Gal Elidan, Hengrui Liu, Holger Winnemoeller, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, and 18 others. 2025. Evaluating gemini in an arena for learning. *arXiv preprint arXiv:2505.24477*.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Lisa Wang, Markus Kunesch, Miruna Pîslar, and 26 others. 2024. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2024. Analyzing large language models for classroom discussion assessment. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024)*, pages 500–510, Atlanta, Georgia, USA. International Educational Data Mining Society.

Jiani Wang, Kamil Hankour, Yuqi Zhang, Jennifer LoCasale-Crouch, and Jacob Whitehill. 2025. Classroom observation evaluation with large language models. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 83–93. PMLR.

Xiaolong Wang, Zhe Zhao, Song Lai, Chaoli Zhang, Zijie Geng, Yu Tong, Ye Wei, and Qingsong Wen. 2026. From memorization to creation: Evaluating the cognitive depth of llm?generated educational questions. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '26, page 2818–2829, New York, NY, USA. Association for Computing Machinery.

Shou'ang Wei, Xinyun Wang, Shuzhen Bi, Jian Chen, Ruijia Li, Jiang Bo, Xin Lin, Min Zhang, Yu Song, BingDong Li, Aimin Zhou, and Hao Hao. 2025. Elmes: An automated framework for evaluating large language models in educational scenarios. *arXiv preprint arXiv:2507.22947*.

Daniel Weitekamp, Momin N. Siddiqui, and Christopher J. MacLellan. 2025. Tutorgym: A testbed for evaluating ai agents as tutors and students. *arXiv preprint arXiv:2505.01563*.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*, 16(1):34–60.

Eamon Worden, Cristina Heffernan, Neil T. Heffernan, and Shashank Sonkar. 2026. Foundationalassist: An educational dataset for foundational knowledge tracing and pedagogical grounding of llms. *arXiv preprint arXiv:2602.00070*.

Wenjing Xie, Juxin Niu, Chun Jason Xue, and Nan Guan. 2025. Grade like a human: Rethinking automated assessment with large language models. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems (RACS 2025)*, pages 21:1–21:8. ACM.

Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Zhuangzhi Dong, Jingren Zhang, Yufan Deng, Xinyu Zou, Yang Gao, and Heyan Huang. 2025. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *arXiv preprint arXiv:2505.16160*.

Tengchao Yang, Sichen Guo, Mengzhao Jia, Jiaming Su, Yuanyang Liu, Zhihan Zhang, and Meng Jiang. 2025. Mmtutorbench: The first multimodal benchmark for ai math tutoring. *arXiv preprint arXiv:2510.23477*.

Bo Yuan and Jiazi Hu. 2024. An exploration of higher education course evaluation by large language models. *arXiv preprint arXiv:2411.02455*.

Mike Zhang, Euan D. Lindsay, Frederik Bode Thorbensen, Danny Bøgsted Poulsen, and Johannes Bjerva. 2024. Leveraging large language models for actionable course evaluation student feedback to lecturers. *arXiv preprint arXiv:2407.01274*.

Min Zhang, Hao Chen, Hao Chen, Wenqi Zhang, Didi Zhu, Xin Lin, Bo Jiang, Aimin Zhou, Fei Wu, and Kun Kuang. 2025. Omniedubench: A comprehensive chinese benchmark for evaluating large language models in education. *arXiv preprint arXiv:2510.26422*.

Chengliang Zhou, Mei Wang, Ting Zhang, Qiannan Zhu, Jian Li, and Hua Huang. 2025a. From answers to questions: Eqgbench for evaluating llms' educational question generation. *arXiv preprint arXiv:2508.10005*.

Yuhang Zhou, Xutian Chen, Yixin Cao, Yuchen Ni, Yu He, Siyu Tian, Xiang Liu, Jian Zhang, Chuanjun Ji, Guangnan Ye, and Xipeng Qiu. 2025b. Teach2eval: An indirect evaluation method for llm by judging how it teaches. *arXiv preprint arXiv:2505.12259*.

Maria-Eleni Zoumpoulidi, Eleni Batsi, Georgios Paraskevopoulos, Vassilis Katsouros, and Alexandros Potamianos. 2025. Bloomxplain: A framework and benchmark dataset for pedagogically sound llm-generated explanations based on bloom's taxonomy. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. Metadata completed from NeurIPS 2025 virtual program and OpenReview record.