The Standard of Proof for AI in EdTech

Applying Measurement Science to AI Use in Education Assessments

Nov 12, 2026

Authors: Alexis Andres, Ed.D., Learning Data Insights, Melanie Kurmichak, Learning Data Insights, Amy Burkhardt, Ph.D., Learning Data Insights, Maggie Beiting-Parish, Ph.D, City University of New York, Michael Feldstein, 1EdTech

AI Disclosure: Claude Sonnet was used for initial drafting of some text, which was edited and reviewed by humans. All concepts and final text have been reviewed and are supported by the authors.

License Type CC BY 4.0 DEED



The question everyone seems to be asking is, “Does AI improve education?” But that question is unanswerable. “AI” is not an intervention. We can’t evaluate the effectiveness of “using AI,” unmoored from a specific educational task and context. We don’t ask, “Do word processors improve education?” And yet, distressingly often, even in scholarly research, the question that gets posed is something like, “Did ChatGPT improve student success?”

The stakes are high. Schools are becoming increasingly concerned about when technology is helpful or harmful in classrooms, and in many cases are forced to make stark choices such as restricting or prohibiting its use because they lack the information they need to make more discriminating decisions about safe and effective uses of EdTech. This is a general problem caused by our lack of clear decision-supporting research, which AI further complicates with its unprecedented and poorly understood capabilities and limitations in educational contexts.

This report attempts to provide clear and practical guidance based on answerable questions. It describes the initial findings of a structured literature review covering three educational assessment tasks: automated scoring, item generation, and formative feedback. We hope the findings that readers find surprising will spark practical conversations about the standard the sector should hold itself to when making claims about what works.

For example, AI can be more trustworthy for automated scoring than for formative feedback, not because we know AI is better at scoring than at providing feedback, but because scoring has a trustworthy way to check answers while formative feedback doesn’t yet. When we have more trustworthy methods, we can have more confidence in counterintuitive results, such as our finding that small, fine-tuned models can be more effective than prompting large frontier models. This finding has implications for access and trust, since a smaller model is less costly to run and offers institutions more potential choice of providers that they trust with their students’ data. Our ability to accept it with confidence is consequential.

The insights provided here are drawn from substantial analysis. Eight researchers from Learning Data Insights and research universities evaluated 250 studies that met the team's quality requirements, from an initial pool of over 3,000 papers, to identify the methods being used to evaluate AI and patterns in the results. This document synthesizes the approaches and methods used in those studies.

All papers are available at https://www.aievidencehub.org/. While the research team plans to publish a full scholarly analysis, Learning Data Insights and 1EdTech are pleased to present this overview of what was found, what was missing, and what was learned.

Educational technology improves when we act collectively as a sector. We can start by asking better questions. What is the intervention that is claimed to "work"? Does it have a trustworthy measure of success? What specific students does it work for? Does the AI consistently give the same answer to the same input? When we decide together on the standards of proof, the sector can move from shutting down technology use out of necessary caution toward adopting products that support safe and effective educational practices.

Results by Application Area

We start with results by each of the areas studied in this analysis: automated scoring, automated item generation, and formative feedback. Each is a different kind of educational intervention that raises different questions and requires different standards of proof.

Automated Scoring

WHY IS MATTERS	FINDINGS	WHAT'S NEXT
AI scoring could cut cost and turnaround, but current fine-tuning per question (2,000+ scored responses) is too expensive for most classroom use.	**Theme 1:** Smaller fine-tuned models beat larger prompted frontier models. **Theme 2:** Rubrics, anchors, and item context improved results more than changing models. **Theme 3:** Human raters often disagree, So AI inherits their inconsistency; check human reliability first. **Theme 4:** AI handles clear cases well; humans should handle ambiguous ones. **Theme 5:** Extra prompting or reasoning steps can hurt accuracy.	**Implications** Report QWK and disaggregated results, supply rubrics, rerun responses, route uncertain cases to humans. **Gaps** Almost no studies test score consistency across runs or integration with teacher judgment.

Why automated scoring matters

Automated scoring assigns a score to a complex student work product that would otherwise be evaluated by a human: an essay, a short constructed response, an interview transcript, or another complex response. When automated systems are as accurate as trained human raters, they can increase the use of innovative assessments, return results faster, and reduce the cost of human scoring.

Generative AI has raised the possibility that general-purpose models could score new responses using only a rubric and a small number of examples, while also providing explanations that teachers can check. If that promise holds, automated scoring could become easier to incorporate into educational products. Automated scoring is frequently used in high-stakes standardized testing. Current standard practice is to fine-tune a transformer model for each individual question, often requiring more than 2,000 human-scored responses. Because this process is costly and item-specific, it is generally not widely used in the classroom or in education technology products. While we found that accuracy is improving with prompt-based models, it still falls short of the threshold for useful classroom results without fine-tuning or other advanced techniques that are out of reach for classroom teachers and education technology providers.

What strong evidence looks like

Automated scoring has a clear evaluation standard because researchers can compare AI scores with human scores. Of the 105 scoring papers we reviewed, 87 (83%) compared AI scores with human scores; of those, 44 (42%) used quadratic weighted kappa (QWK), the standard metric for agreement on ordinal scales (e.g., scores ordered from low to high). Established evaluation standards require AI systems to perform at a level comparable to human raters, without systematic score differences or large performance gaps for any student groups. Only 13% of scoring papers relied on custom metrics, such as aptness index (Gandolfi, 2025) and different difference thresholds (Mendonça et al., 2025), demonstrating that automated scoring research more often uses established evaluation approaches than the other domains we reviewed.

What we found

Theme 1: Training approach mattered more than model size

Fine-tuned, smaller models outperformed frontier models. Zero-shot and few-shot prompting of frontier models (GPT-3.5, GPT-4, Llama, Mistral) produced lower agreement than models fine-tuned on scored responses in standard essay scoring benchmarks. In one study, a 37-million-parameter encoder reached a trait-level QWK of 0.70, outperforming a model twenty times its size as well as GPT-4 prompted directly (Han et al., 2025). Across studies, evaluation approach and training methods appeared to matter more than model family alone. For many product and engineering teams, this result runs against the common intuition that a bigger model is always better. Smaller models that are fine-tuned on scored student responses matched or outperformed much larger models that were only prompted (Chamieh et al., 2024; Hou et al., 2025; Mansour et al., 2024; Ormerod & Kwako, 2024; Wang et al., 2024; Xiao et al., 2024). However, fine-tuning requires more up-front investment,

including collecting and scoring student responses. Once that work is done, a smaller model can be cheaper to run at scale, may be easier to adapt to new questions than traditional (pre-LLM) automated scoring engines, and gives teams more control over student data governance and deployment. This situation creates an opportunity for openly released, moderately sized models fine-tuned for assessment tasks to make high-quality automated scoring more accessible. So far, the evidence comes mostly from scoring tasks with well-defined rubrics and enough scored examples to train on; future research should test whether this pattern holds for items without these materials.

Theme 2: Item context materials matter more than model choice
Beyond fine-tuning with student responses, factors such as rubrics, training materials, and item context strongly affected results. Adding information about the scoring task improved performance *more than changing the underlying AI model*. For example, one study found that adding a one-sentence item description to each training response moved models from failing every evaluation threshold (QWK 0.00–0.75) to near-human agreement (0.68–0.93) (Jin et al., 2025). The results suggest that giving models better examples and context may be more valuable than relying on more sophisticated prompting strategies. This result was truer for Automated Scoring than in other domains because 'anchor papers' representing each scoring band are often available.

Theme 3: Human agreement sets the standard for AI evaluation
Human scoring variability shapes how we should evaluate automated scoring systems. It provides a clear comparison case and sets the limits for what we can reasonably expect from AI. In one study, two teachers scoring the same math answer agreed barely above chance. An AI system trained on a single teacher's scores would inherit that teacher's scoring habits (Zhang et al., 2023). This can lead to unreliable scores, bias, and fairness concerns. As such, when evaluating an automated scorer against a human scorer, it is important to first consider the consistency of human scoring.

Theme 4: Human and AI systems can complement each other
Combining human reviewers and AI systems can leverage their different strengths and limitations, improving performance by allowing AI systems to handle clearer cases, while educators focus on addressing ambiguous responses from students that require more judgement (Grévisse, 2024). AI systems also identified 'top entries' in a competition with near-perfect accuracy, while humans selected the final winner from a much smaller set of entries (Cui & Liang, 2024).

Theme 5: More prompting does not always improve results
Clear instructions about the scoring task matter, but adding more prompting or reasoning steps does not necessarily improve performance. In one study, using Chain of Thought reasoning (CoT) to correct an error caused the model to mislabel other responses it had previously scored correctly (Cohn et al., 2024). Another study found that GPT-3.5 performance decreased as the number of examples in the prompt increased (Chamieh et al., 2024).

IMPLICATIONS FOR PRODUCT TEAMS		
COMPARE Score student responses using humans, then compare AI scores. Report QWK, raw agreement, and mean-score difference.	**REPORT DISAGGREGATED RESULTS** Report results by score level and learner subgroup, not as one overall number.	**SUPPLY AI WITH RUBRICS** Provide the full rubric and at least one anchor response per score point. Test prompt changes on varied held-out data.
RERUN RESPONSES Run the same responses multiple times. Determine if scores change.	**ROUTE TO HUMANS** Send low-confidence and mid-range responses to human review. Report the share that needs review.	DO NOT Treat agreement with another automated system, or results on synthetic data, as evidence.

Remaining evidence gaps

The research leaves several important questions open.

1. How consistently will AI score the same student work? Almost no studies report what happens when the same response is scored more than once, so we don't know how stable most published results are. In practice, if a student could get a different score by chance, the system is unreliable.

2. How can AI scoring enhance teachers' understanding of students? The intent of AI scoring is to provide students with more feedback on their work and opportunities to practice, and not to replace teacher judgment. How can we integrate these systems with teacher assessments offered in classrooms?

Answering these questions will be essential to ensuring automated scoring improves both the quality and frequency of evaluation.

Automated Item Generation

WHY IS MATTERS	FINDINGS	WHAT'S NEXT
LLMs could generate items at scale and put item writing in more hands, but that also removes expert review.	Theme 1: Generated items nearly always need human editing, some have wrong answer keys, Theme 2: Context and grounding matter more than model size or family. Theme 3: Quality varies by item type, subject, and format; no model wins everywhere.	**Implications** Review every item, verify keys, give context, pilot with real students, check fairness. **Gaps** Few studies pilot at scale, test fairness, or measure learning; unclear whether non-experts can match researchers' results.

Why automated item generation matters

Automated item generation (AIG) uses AI to create assessment content, including multiple-choice items with plausible distractors, reading passages and comprehension questions, and more. Creating good assessment items is difficult. Generating items on demand with trained Large Language Models (LLMs) could produce more test items, faster content refreshes, and practice material tailored to a particular standard, text, or student. LLMs enable large-scale item development across subjects, with items that can resemble human-written content, including distractors based on common student misconceptions, and adjustable difficulty for different users and ability levels. This approach also opens the door to democratization. If good items can be generated without expert prompting, instructors with deep subject knowledge but little psychometric training could produce sound test questions by drawing on their domain expertise and the context of their own course. The promise and the peril here emerge from the ease of use that puts item writing in more hands and also removes the expert review that has traditionally caught flawed items.

The research studies reviewed largely use three approaches to establish the validity of test items. First, evaluative review relies on evaluations of the generated items (e.g., Likert-scale ratings, text complexity scores), which humans or models can perform. Second, automated text similarity comparisons assess how generated items compare with existing items or similar text. Finally, the most robust studies combined these approaches with Pilot testing on students to examine real-world performance. Researchers can use these approaches individually or in combination.

What we found

Theme 1: Human review remains necessary

AI-generated items still require human review before use. The amount of editing needed varies by study, but generated items are rarely ready to use as-is. Even the strongest of five frontier models required edits on 78% of items (Zeinalipour et al., 2025). Another study found that 85% of generated items were usable with minor edits (Li et al., 2025). In a separate evaluation, only 32% of ChatGPT 3.5 practice questions were usable as written and 25% had wrong answers marked correct (Ngo et al., 2023).

Theme 2: Task context and grounding can matter more than model choice

Model size and family were not the determining factors affecting performance. A smaller language-specific model outperformed general LLMs in one study (Awalurahman & Budi, 2026). In another, adding a clinical-history requirement to the prompt improved the quality of medical items (Klang et al., 2023).

Theme 3: Performance varies by task and content area

No single model performed best across every type of item. Results also varied by passage type. In one study, AI-generated literary passages were rated as more engaging and easier to comprehend than the human-written comparison passages, while informational passages showed more challenges with organization and coherence than human-authored items (Bezirhan & von Davier, 2023). In medical item generation, error rates varied by specialty, with surgery items showing more errors than psychiatry items (Klang et al., 2023).

Model performance also varied by task format, with near-ideal performance on binary and multiple-choice tasks but substantially weaker performance on causal, counterfactual, and close tasks (Zeinalipour et al., 2025). Character count alone predicted item difficulty better than an LLM simulator in one study (Ma et al., 2025), and image-based questions remained more challenging than text-based items (Klang et al., 2023). One study also found that reviewers could not reliably distinguish AI-generated from human-written informative texts, while human-written narratives and other creative writing tasks remained higher quality (Ripoll Y Schmitz & Sonnleitner, 2025).

IMPLICATIONS FOR PRODUCT TEAMS		
REVIEW Have a human review every generated item before use.	**CHECK ANSWER KEYS** Verify marked answers and explanations.	**GIVE CONTEXT** Give the model context, such as source text or required clinical history.
REPORT QUALITY Report quality by item type, content area, and passage type.	**PILOT ON REAL STUDENTS** Administer items to real students. Report psychometric properties next to human-written items.	**CHECK FAIRNESS** Check whether generated items perform fairly across learner subgroups.

Remaining evidence gaps

1. Few studies have administered AI-generated items to real students at scale and reported psychometric properties alongside human-written items. Fewer studies have tested whether generated items are fair across subgroups or whether AI-generated practice items change what students learn and how well they perform. The individual approaches for evaluating item quality are well established, but the field has not yet agreed which combination of evidence is sufficient for different uses. For consequential applications, developers should supply evidence across all three approaches.

2. The research also does not tell us how much expertise is needed to get good results. The studies reviewed here were run by researchers who wrote and refined their prompts with care, and the evidence shows that item quality depends heavily on the context and grounding the model is given. Whether a teacher without psychometric training, working from course materials alone, or with the help of a prebuilt prompt template, can reach the same quality is an open question that these studies do not directly answer.

Formative Feedback

WHY IS MATTERS	FINDINGS	WHAT'S NEXT
Feedback is a powerful but scarce lever; AI could expand access, but must first correctly diagnose the error.	**Theme 1:** AI feedback rarely outperforms human feedback on learning outcomes. **Theme 2:** Students often like AI feedback with learning more from it. **Theme 3:** LLMs default to giving answers instead of hints and struggle to diagnose answers. **Theme 4:** Accuracy is rarely measured and varies widely by task.	**Implications** Review and audit feedback, verify instruction-following, design for productive struggle, measure learning, disclose methods. **Gaps** Few rigorous learning studies (small or null-effects), ~20% error rates go unnoticed, fairness and K-12 evidence are thin, cost of human review unreported.

Why formative feedback matters

Formative feedback gives students information about their work that they can use to improve: a hint on a wrong math answer, a comment on a draft, a correction to a sentence. Earlier research has established that good feedback helps students understand where they are going, how they are progressing, and what to do next (Hattie & Timperley, 2007). Timely, specific feedback is one of the most reliable levers to improve learning, but it is also one of the scarcest resources in a classroom. GenAI makes immediate, individualized, and unlimited feedback possible. Feedback requires accurately identifying an issue to address, so it is inherently more complicated than the other two domains.

What we found

Theme 1: Evidence of learning impact remains limited

Evidence that AI-generated feedback improves student performance is still limited. Measuring learning effects is difficult in real educational settings, but the findings so far rarely show AI feedback outperforming human feedback. In an undergraduate programming course, students who received instructor feedback improved their revisions more than those who received ChatGPT feedback (Er et al., 2025). In a study with medical students, LLM feedback personalized to each student's response did not lead to higher exam scores compared with a single expert comment given to all students. Scores actually trended slightly in favor of the expert comment (Nissen et al., 2025). Other smaller studies found no significant difference between AI and human feedback on writing revisions (Alnemrat et al., 2025; Escalante et al., 2023). A randomized study found similar performance when a human tutor reviewed and edited every AI-drafted message before students saw it. Students were about as likely to correct a math mistake as those helped by a human tutor alone. However, this study tested an AI system with human oversight rather than autonomous AI feedback, measured results on the same day, and was conducted by the companies that built the tool (LearnLM Team et al., 2025).

Theme 2: Student perceptions are not the same as learning impact

How students perceive feedback matters, but feedback only helps if students read it, take it seriously, and act on it. Positive reactions are only one piece of evidence that feedback is well designed. Students may value feedback that does little to improve their learning. Some studies show a gap between how students rate AI feedback and how much it helps them (e.g., Nissen et al., 2025). In a design course, 25 students rated GPT-4 feedback higher than professors' feedback on every measured dimension, including encouragement and structure, but the study measured no learning improvement, and AI's consistently positive tone may have influenced those ratings (Schmitt-Fumian et al., 2025). When students received both human and AI feedback, preferences split roughly evenly: some valued human tutors for interaction, while others valued AI for clarity and availability (Escalante et al., 2023). Usage also differed from designers' expectations. On one physics problem, all 61 students used "check my work" feedback, but only 9 asked for the next step, and the AI chat tutor was students' least-liked feature (Dange et al., 2026). Together, these findings suggest that student satisfaction is a useful but incomplete signal, and evaluations of AI feedback should pair perception measures with evidence of improved work or learning outcomes.

Theme 3: AI feedback must support student thinking

Effective formative feedback helps students work through problems themselves. Simply telling students the answer may fix the immediate mistake, but it can bypass the reasoning students need to avoid repeating the same error, a pattern called "cognitive offloading" and a major risk of using AI in education. LLMs tend to default to giving answers. In one study, GPT-4o gave hints only 0–3% of the time even when prompted to do so, while 41% for teachers did (Coyne et al., 2025). Even tools designed to encourage self-correction can still supply direct fixes, such as rewritten sentences in grammar feedback (Escalante et al., 2023). Some tools are deliberately designed to avoid simply giving answers. One highlights where an error occurred and asks students to fix it themselves (Young et al., 2025); another deducts points when students use hints or view solutions, encouraging them to check their own work first (Dange et al., 2026). These approaches are promising, but whether they actually preserve productive struggle or simply discourage requests for help remains largely untested. Feedback that avoids giving the answer is only useful if it correctly identifies the problem. Across methods, LLMs most often struggled to diagnose why a student made an error, especially on multi-step problems (Scarlatos et al., 2024).

Theme 4: Feedback accuracy remains difficult to evaluate

Feedback is only helpful when it is correct, yet accuracy is one of the least consistently examined aspects of GenAI feedback. When researchers have measured accuracy, results vary widely by task. In one study, GPT-4o feedback on English learners' grammar errors was rated factually correct about 97% of the time (Coyne et al., 2025), while a commercial writing tool detected some types of errors over 90% of the time and others under 20%, and its suggested corrections matched the correct fix only 17% to 54% of the time (Young et al., 2025). These measures are essential for judging whether a tool can be trusted, yet few studies report them. Without auditing a sample of AI feedback before or during use, it is impossible to know whether weak learning outcomes reflect ineffective or inaccurate feedback.

IMPLICATIONS FOR PRODUCT TEAMS		
REVIEW Use human reviewers and editors for AI-generated feedback before students see it.	**AUDIT** Check a sample of feedback for accuracy, and keep checking over time.	**VERIFY** Test whether the model follows prompt instructions in practice, such as giving hints instead of answers.
DESIGN Build for productive struggle, not just correct answers.	**MEASURE** Track learning outcomes and student perceptions of feedback.	**DISCLOSE** Be transparent about methods when reporting learning outcomes.

Remaining evidence gaps

1. The largest gap is whether AI feedback improves learning. Only 18 of 55 papers measured a learning or behavioral outcome, but only two (Bauer et al., 2025; Kinder et al., 2025) randomized students against a comparison condition a school would plausibly offer and reported effect sizes. Both found small or null effects on performance.

2. Few studies audit whether students receive correct feedback. In the studies that do, error rates run about 20% (Maus et al., 2025; Qian et al., 2025), and students rarely notice. Only 6% of feedback papers tested fairness, one disaggregated by language status (Steiss et al., 2024), none examined students with disabilities, and K-12 evidence rests on a handful of papers. The studies with the strongest safety records had a person review every AI message (LearnLM Team et al., 2025), but none report what that review costs or what happens when it is removed.

Overall Results & Questions to Ask

Across the different application areas, three questions consistently distinguish stronger evaluations from weaker ones:

1. **What is the 'ground truth' to measure AI quality against?** What are the AI results being compared against, and how reliable is that comparison? In automated scoring, the reference should usually be human scores. But the comparison only matters when human raters meet a standard of agreement. If they don't, the construct is either unclear or too complex to score consistently, and an AI score that "matches" a human score isn't necessarily correct.

2. **Are the evaluation metrics comparable or study-specific?** In item generation and feedback, there often is no human reference to check against, because these systems create items and comments that didn't exist before. Sometimes researchers use another model's judgment as the reference. That's a weaker evaluation because AI systems may rate other AI output more favorably than human reviewers. Most studies use custom

evaluations, rubrics, and scoring guidelines designed for their specific context. Interpret these results within each study, not directly across studies or vendors.

3. **Were fairness and reliability tested?** Fairness means checking whether the AI works equally well for different groups of students, for example, whether it scores English learners, students with disabilities, or speakers of different dialects as accurately as other students. Reliability means checking whether the AI gives the same result when it receives the same input more than once. AI models may give different scores or feedback each time the same student response is submitted, so a single run may not reflect how the system behaves in practice. This is obviously a problem if you are creating learning materials or a student is seeking to solve a problem that they are struggling with. If a study does not report results by student subgroup (fairness) and consistency across repeated runs (reliability), treat both as unknown, and don't assume the results will apply to your students.

Guidance for Developing and Testing AI in Education

When evaluating GenAI output, some strategies provide stronger evidence for claims supporting its use, while others threaten validity, reliability, and fairness.

DONT		DO
Assume a model that is lenient, strict, or accurate today will behave the same after a model update.	→	Re-test after every model update, and record which model and prompt version produced each reported result.
Rely on prompt instructions ("give hints, ignore grammar errors") as safeguards.	→	Test whether the model follows instructions; provide ample context and examples.
Start with open-ended applications, such as AI tutoring, when you need a new evidence base.	→	Start with uses where established evaluation standards already exist, such as scoring against human raters.
Treat student satisfaction or preferences as evidence that AI feedback works.	→	Pair perception measures with quantitative evidence of improved student work or learning.
Accept effectiveness claims without considering how the study was designed.	→	Ask what the AI was compared against, who ran the study, and what outcome was measured.
Accept the applicability of an AI tool with a new population	→	Establish fairness checks by exploring the representativeness of the data being used to train and fine-tune models.

What this means in practice

For educators and families:

- AI-assigned scores without a human in the loop can be wrong in hard-to-detect ways.
- Treat AI feedback as a first pass, and teach students to check it against the original source or assignment criteria, or ask their teacher for clarification.
- Review AI-written practice items to confirm that answers and explanations are correct.
- Use additional sources of evidence outside of AI generated/scored assessment materials to ensure valid and reliable measures of student growth/learning
- AI scoring products are rarely evaluated for fairness and bias, so be aware of this when using AI tools for education

For procurement and implementation teams:

- Look for vendors that are transparent about their prompting, training, evaluation practices, and the populations used for training, fine-tuning, and testing.
- Ask which prompt and model version produced reported results, and what happens when models are updated.
- Ask what proportion of outputs are reviewed by humans.
- Ask about the background of the model training/development team: are they well-rounded, or are there gaps missing that may lead to issues with the tool long-term?
- Does the vendor provide any product specifications or benchmark scores to demonstrate product quality? If so, what are these and how are they represented?

For product teams:

- Current practices are more reliable for responses that are clearly correct or clearly incorrect.
- Responses that fall in the middle of a score range still benefit from human judgment.
- Report accuracy by score level and by learner subgroup.
- Supply the full rubric and at least one anchor response for each score point.
- Explore the optimal number of work samples to most accurately train your product for its intended purpose.
- Use a defensible and psychometrically valid benchmark for ensuring external model quality.

References

Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., & AlSharefeen, R. (2025). AI vs. teacher feedback on EFL argumentative writing: A quantitative study. *Frontiers in Education, 10*, Article 1614673. https://doi.org/10.3389/feduc.2025.1614673

Awalurahman, H. W., & Budi, I. (2026). Multi-task fine-tuning using prefix-prepend format in text-to-text transfer transformer for automatic question generation in Bahasa Indonesia. *PeerJ Computer Science*, 12, Article e3987. https://doi.org/10.7717/peerj-cs.3987

Bauer, E., Richters, C., Pickal, A. J., Klippert, M., Sailer, M., & Stadler, M. (2025). Effects of AI-generated adaptive feedback on statistical skills and interest in statistics: A field experiment. *British Journal of Educational Technology, 56*(5), 1735–1757. https://doi.org/10.1111/bjet.13609

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence, 5*, Article 100161. https://doi.org/10.1016/j.caeai.2023.100161

Chamieh, I., Zesch, T., & Giebermann, K. (2024). LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 309–315). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.25/

Cohn, C., Hutchins, N., Le, T., & Biswas, G. (2024). A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(21), 23182–23190. https://doi.org/10.1609/aaai.v38i21.30364

Coyne, S., Galvan-Sosa, D., Spring, R., Guerraoui, C., Zock, M., Sakaguchi, K., & Inui, K. (2025). Annotating errors in English learners' written language production: Advancing automated written feedback systems. In *Artificial Intelligence in Education: 26th International Conference, AIED 2025* (Lecture Notes in Computer Science, Vol. 15880, pp. 292–306). Springer. https://doi.org/10.1007/978-3-031-98459-4_21

Cui, Y., & Liang, M. (2024). Automated scoring of translations with BERT models: Chinese and English language case study. *Applied Sciences, 14*(5), Article 1925. https://doi.org/10.3390/app14051925

Dange, A., Lopez, R. E., Shah, N., & Deslauriers, L. (2026). *aiPlato: A novel AI tutoring and stepwise feedback system for physics homework*. arXiv. https://arxiv.org/abs/2601.09965

Er, E., Akçapınar, G., Bayazıt, A., Noroozi, O., & Banihashem, S. K. (2025). Assessing student perceptions and use of instructor versus AI-generated feedback. *British Journal of Educational Technology, 56*(3), 1074–1091. https://doi.org/10.1111/bjet.13558

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, Article 57. https://doi.org/10.1186/s41239-023-00425-2

Gandolfi, A. (2025). GPT-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions. *International Journal of Artificial Intelligence in Education, 35*(1), 367–397. https://doi.org/10.1007/s40593-024-00403-3

Grévisse, C. (2024). LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education, 24*, Article 1060. https://doi.org/10.1186/s12909-024-06026-5

Han, D., Roh, D., Han, E., Song, H., & Yi, M. Y. (2025). Fine-grained multi-prompt essay scoring with multi-level disentanglement. *Data Mining and Knowledge Discovery, 39*(5), Article 67. https://doi.org/10.1007/s10618-025-01126-2

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

Hou, Z., Ciuba, A., & Li, X. (2025). Improving LLM-based automatic essay scoring with linguistic features. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop* (Proceedings of Machine Learning Research, Vol. 273, pp. 41–65). PMLR. https://proceedings.mlr.press/v273/hou25a.html

Jin, H., Lima, C., & Wang, L. (2025). Automated scoring in learning progression-based assessment: A comparison of researcher and machine interpretations. *Educational Measurement: Issues and Practice, 44*(3), 25–37. https://doi.org/10.1111/emip.70003

Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers and Education: Artificial Intelligence, 8*, Article 100349. https://doi.org/10.1016/j.caeai.2024.100349

Klang, E., Portugez, S., Gross, R., Kassif Lerner, R., Brenner, A., Gilboa, M., Ortal, T., Ron, S., Robinzon, V., Meiri, H., & Segal, G. (2023). Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: A medical education pilot study with GPT-4. *BMC Medical Education, 23*, Article 772. https://doi.org/10.1186/s12909-023-04752-w

LearnLM Team, Google, & Eedi. (2025). *AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms*. Google DeepMind. https://storage.googleapis.com/deepmind-media/LearnLM/learnLM_nov25.pdf

Li, J., Rezayi, S., Baldwin, P., Harik, P., & Yaneva, V. (2025). Towards reliable generation of clinical chart items: A counterfactual reasoning approach with large language models. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers* (pp. 142–153). National Council on Measurement in Education. https://aclanthology.org/2025.aimecon-main.16/

Ma, W. A., Flor, M., & Wang, Z. (2025). Automatic generation of inference making questions for reading comprehension assessments. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)* (pp. 398–414). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.bea-1.31

Mansour, W. A., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2777–2786). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.247/

Maus, H., Tschisgale, P., Kieser, F., Petersen, S., & Wulff, P. (2025). *Developing and evaluating a large language model-based automated feedback system grounded in evidence-centered design*. arXiv. https://arxiv.org/abs/2512.10785

Mendonça, P. C., Quintal, F., & Mendonça, F. (2025). Evaluating LLMs for automated scoring in formative assessments. *Applied Sciences, 15*(5), Article 2787. https://doi.org/10.3390/app15052787

Ngo, A., Gupta, S., Perrine, O., Reddy, R., Ershadi, S., & Remick, D. (2023). ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Academic Pathology, 10*(4), Article 100099. https://doi.org/10.1016/j.acpath.2023.100099

Nissen, L., Rother, J. F., Heinemann, M., Reimer, L. M., Jonas, S., & Raupach, T. (2025). A randomised cross-over trial assessing the impact of AI-generated individual feedback on written online assignments for medical students. *Medical Teacher, 47*(9), 1544–1550. https://doi.org/10.1080/0142159X.2025.2451870

Ormerod, C., & Kwako, A. (2024). *Automated text scoring in the age of generative AI for the GPU-poor*. arXiv. https://arxiv.org/abs/2407.01873

Qian, K., Cheng, Y., Guan, R., Dai, W., Jin, F., Yang, K., Nawaz, S., Swiecki, Z., Chen, G., Yan, L., & Gašević, D. (2025). *Dean of LLM tutors: Exploring comprehensive and automated evaluation of LLM-generated educational feedback via LLM feedback evaluators*. arXiv. https://arxiv.org/abs/2508.05952

Ripoll Y Schmitz, L. M., & Sonnleitner, P. (2025). Evaluating AI-generated vs. human-written reading comprehension passages: An expert SWOT analysis and comparative study for an educational large-scale assessment. *Large-scale Assessments in Education, 13*, Article 20. https://doi.org/10.1186/s40536-025-00255-w

Scarlatos, A., Smith, D., Woodhead, S., & Lan, A. (2024). Improving the validity of automatically generated feedback via reinforcement learning. In *Artificial Intelligence in Education: 25th International Conference, AIED 2024* (Lecture Notes in Computer Science, Vol. 14829, pp. 280–294). Springer.

Schmitt-Fumian, T. K., Tauscher, S., & Thoring, K. (2025). AI vs. human: Exploring the potential of generative AI as a feedback tool to support ideation in design education. *Proceedings of the Design Society, 5*, https://doi.org/10.1017/pds.2025.10057

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Wang, Y., Hu, R., & Zhao, Z. (2024). Beyond agreement: Diagnosing the rationale alignment of automated essay scoring methods based on linguistically-informed counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 8906–8925). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.520

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). *Human-AI collaborative essay scoring: A dual-process framework with LLMs*. arXiv. https://arxiv.org/abs/2401.06431

Young, M., Burkhardt, A., Wilson, Q., & Lottridge, S. (2025, April). *A framework for evaluating automated conventions feedback* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, Denver, CO, United States. https://files.portal.cambiumast.com/corporate-site/documents/CAI-A-Framework-for-Evaluating-Conventions.pdf

Zeinalipour, K., Jamshidi, N., Hejazi, S., Maggini, M., Bianchini, M., Paoletti, S., & Gori, M. (2025). FarSense: A comprehensive commonsense benchmark and evaluation framework for the Farsi language. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (pp. 3529–3599). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.ijcnlp-long.187

Zhang, M., Heffernan, N., & Lan, A. (2023). Modeling and analyzing scorer preferences in short-answer math questions. In *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 378–387). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115796