Defining Quality: Teacher Priorities for Evaluating AI-Generated Lesson Plans

AI in the Classroom: Benchmarking and Testbed Framework for Evidence-Based Adoption

Ruhan Circi, Bhashithe Abeysinghe, Kelsey Woodrick
American Institutes for Research

Introduction

Artificial intelligence (AI) is quickly reshaping education. There are many ways teachers may incorporate AI into their classrooms, including using AI tools to help plan instruction. However, what does a high-quality AI-generated lesson plan look like? That question sits at the center of any effort to benchmark or evaluate AI tools. A benchmark is only as meaningful as the quality criteria against which it is measured.

This brief presents findings from a survey of 48 Grade 3–5 teachers about how they evaluate lesson plans that were created with artificial intelligence (AI). It is the second brief in the American Institutes for Research (AIR) series, *AI in the Classroom: Benchmarking and Testbed Framework for Evidence-Based Adoption in K–12.* In the first brief, we examined the landscape of AI tools for K–12 teachers and found substantial variation in the nature and capabilities of the tools.

About This Brief

This brief is part of a series that analyzes the rapidly evolving AI tool environment and introduces systematic, transparent, and relevant approaches to evaluating AI tools in education.

The next brief will provide more information about how surveyed teachers use AI in their classrooms.

What We Know Already

- Teachers adopt AI unevenly across tasks, and their comfort in using AI is shaped by how much professional judgment a task requires.
- Lesson planning is one task where adoption is the strongest, making it a well-supported first target for benchmarking.

What Is New

- Teachers favor a small set of criteria to judge the quality of AI-generated lesson plans. These criteria include alignment to standards (79%), accuracy of content (54%), and appropriateness for grade level (44%).

- Teachers also rely on their expertise to judge AI-generated lesson plans. Only 6 of 48 teachers who completed our survey reported that they use AI plans mostly as is. Most teachers make small edits to the AI lesson plans, significantly adapt them, or use the AI-generated plans to brainstorm ideas.

Why Quality Criteria Matter for Benchmarking

Benchmarks do more than measure performance. They use task definitions, data sets, metrics, aggregation methods, and documentation to operationalize a specific idea of what constitutes a "good" or "high-quality" output. As a result, benchmarks reward systems that optimize the criteria specified for "high-quality" output, even when the resulting outputs may not be useful in practice (Alzahrani et al., 2024; Bowman & Dahl, 2021; Liu et al., 2024; Wallach et al., 2025). For AI-generated lesson plans, benchmark criteria should therefore reflect both research-based evidence and the professional judgment of practicing teachers. Research on the use of AI in education suggests that large language models can support content generation and instructional planning. However, using AI effectively in classroom settings depends on more than producing coherent text. Teachers must apply their professional judgment, pedagogical expertise, and human oversight to ensure that AI-generated materials meet instructional goals as well as student needs.

Lesson planning illustrates this distinction clearly. A lesson plan may be coherent, accurate, and well structured, but it will be deemed as instructionally weak if it:

- Does not align with standards;
- Is not appropriate for a specific grade level;
- Lacks differentiation, meaning instruction is not adapted to students' varying readiness levels, language proficiency, interests, and learning needs to support access to grade-level literacy instruction (Tomlinson et al., 2003); and
- Cannot be feasibly implemented within the constraints of a typical classroom, including available instructional time, resources, and student needs.

Thus, understanding which of these qualities are prioritized by teachers is essential for building benchmarks that measure not only the technical quality of AI-generated materials but also how useful the materials will be in the classroom.

What Questions Did We Ask Teachers?

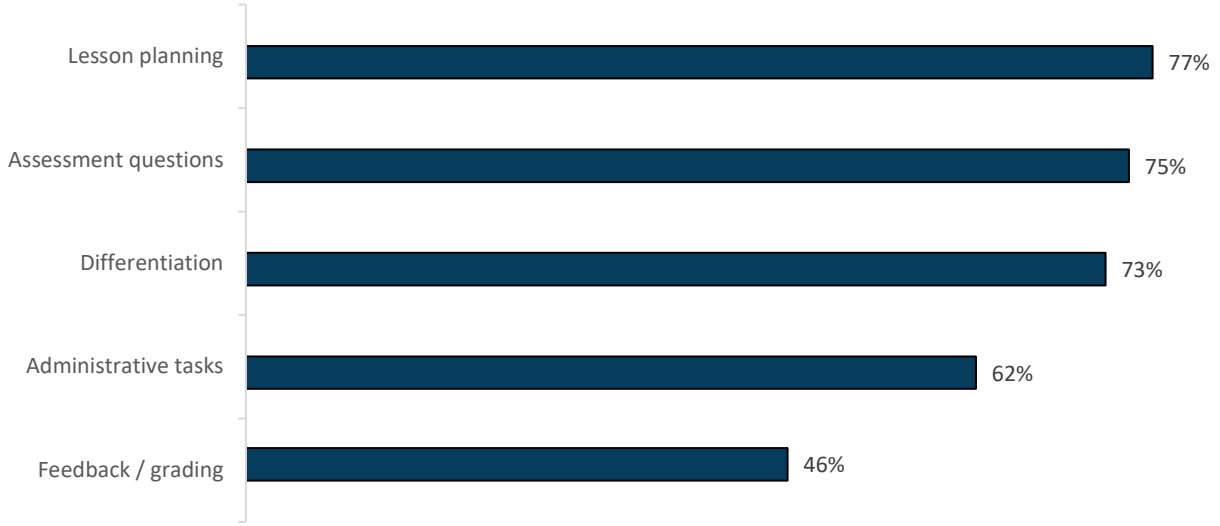
The survey collected information about teaching context, AI use, and comfort using AI for instructional work. One section of the survey was focused exclusively on lesson planning. A total of 48 teachers reported whether they used AI for lesson planning; how they used it (e.g., to generate full plans, to generate lesson components, to adapt existing lessons, or to brainstorm ideas); how much they modified AI-generated content; and which criteria mattered most when evaluating the quality of an AI-generated lesson plan.

What Insights Did We Gain?

1. Teachers Show Strong Comfort with AI-Supported Lesson Planning

Teacher comfort with AI use is highest for planning-related work and lowest for judgment-heavy tasks. For this finding, "comfortable" combines teachers who reported being either very comfortable or somewhat comfortable using AI for each task. Roughly three quarters of teachers are comfortable using AI for lesson planning (77%), to generate assessment questions (75%) and for differentiation (73%). In addition, 62% of teachers report that they are comfortable using AI for planning and preparation. Teacher comfort with AI falls sharply for feedback and grading (46%); this is the task that teachers most often reserve for themselves.

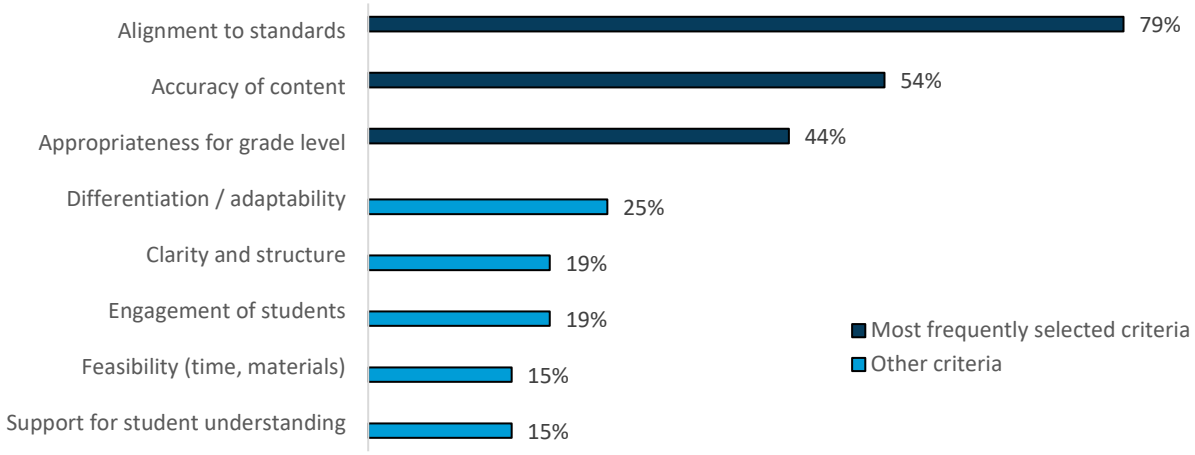
Exhibit 1. Percentage of Teachers Who Reported Feeling Comfortable Using AI, by Task



2. Teachers Prioritize Standards Alignment When Evaluating AI-Generated Lesson Plans

We asked teachers to select up to three criteria that matter most when they are evaluating an AI-generated lesson plan. Seventy-nine percent of teachers selected alignment to standards, making it the most frequently identified criterion. Accuracy of content (54%) and appropriateness for grade level (44%) were the next most frequently selected criteria. A second tier of priorities emerged, with teachers also emphasizing differentiation, clarity, student engagement, feasibility, and support for student understanding.

Exhibit 2. Criteria That Matter Most to Teachers When Judging AI-Generated Lesson Plans



3. Teachers Retain Control When Using AI for Lesson Planning

Two thirds of teachers said they use AI at least occasionally for lesson planning. But they overwhelmingly treat it as a drafting partner rather than an author. When asked what they do with AI, teachers were more likely to report using it to create parts of a lesson or to brainstorm ideas and a plan structure than to adapt an existing or create a complete lesson plan. When asked how much they modified what AI produces, only a few teachers said that they use plans mostly as is. The rest of the teachers reported that they made small edits, significantly adapted the output, or used the output for idea generation.

Exhibit 3. Teachers' Uses of AI-Generated Lesson Planning Support

How teachers work with AI lesson plans	Percentage of teachers (%)
Create parts of a lesson (e.g., activities, questions)	60
Brainstorm ideas or lesson structure	52
Adapt an existing lesson	35
Create a full lesson plan	17

From Teacher Criteria to Benchmark Dimensions

These findings should be viewed as exploratory. The analysis is based on a small sample of Grade 3-5 teachers and is not intended to represent the broader teaching population. However, the survey provides an initial, teacher-informed foundation for developing evaluation dimensions for a Grade 4 reading lesson-planning benchmark, which is the focus of our current initiative. The criteria teachers prioritized can be translated into measurable benchmark dimensions.

The findings have two implications for benchmark design. First, a benchmark should evaluate not only technical accuracy but also the contextual quality of lesson plans, including standards alignment, grade-level appropriateness, differentiation, and overall fit for the intended instructional context. Second, the benchmark should reflect authentic patterns of AI use. Because teachers rarely adopt AI-generated lesson plans without modification, evaluation tasks that focus on generating editable lesson components aligned to specific standards and grade levels are likely to provide a more meaningful measure of classroom utility than tasks that assess fully developed lesson plans in isolation.

Suggested Citation

Circi, R., Abeysinghe, B., Woodrick, K. (2026). *Teacher priorities for evaluating AI-generated lesson plans*. American Institutes for Research.

References

Alzahrani, N., Alyahya, H., Alnumay, Y., AlRashed, S., Alsubaie, S., Almushayqih, Y., Mirza, F., Alotaibi, N., Al-Twairesh, N., Alowisheq, A., Saiful Bari, M., & Khan, H. (2024). When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers) (pp. 13787-13805). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.744

Bowman, S. R., & Dahl, G. (2021). What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4843–4855). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.385

Liu, Y. L., Blodgett, S. L., Cheung, J. C. K., Liao, Q. V., Olteanu, A., & Xiao, Z. (2024). ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers) (pp. 16349–16365). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.861

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., Conover, L. A., & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*(2-3), 119-145.

Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Pangakis, N., Reed, S., Sheng, E., Vann, D., Wortman Vaughan, J., Vogel, M., Washington, H., & Jacobs, A. Z. (2025). Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. https://proceedings.mlr.press/v267/wallach25a.html

Copyright © 2026 American Institutes for Research®. All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, website display, or other electronic or mechanical methods, without the prior written permission of the American Institutes for Research. For permission requests, please use the Contact Us form on AIR.ORG.

Notice of Trademark: "American Institutes for Research" and "AIR" are registered trademarks. All other brand, product, or company names are trademarks or registered trademarks of their respective owners.

31630_10/26