What's Out There? Key Insights From a Scan of AI Tools for K–12 Teachers

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

AIR® Advancing Evidence. Improving Lives.

Ruhan Circi; Bhashithe Abeysinghe; Sara Sharick; Burhan Ogut; Bobbi Newman; Kelsey Woodrick
American Institutes for Research

Introduction

Teachers nationwide are experimenting with a wave of artificial intelligence (AI) tools designed to aid in lesson planning, instructional materials development, grading, classroom management, and many more functions. Understanding what these tools promise and how they function in practice is both essential and underexplored.

To address this gap, AIR's *AI in the Classroom: Benchmarking and Testbed Framework for Evidence-Based Adoption in K–12* project conducted a landscape scan of teacher-facing AI tools currently available in K–12 education. *AI in the Classroom* is part of the broader *AI in Education Network*, which aims to answer the question: What works and what doesn't for AI in classrooms? Through this process, our team at AIR aims to provide educators, district leaders, and developers with evidence-based insights.

About This Brief

This brief examines the current landscape of artificial intelligence (AI) tools used by K–12 teachers, drawing on an initial landscape review conducted by AIR.

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 outputs in education.

The next brief will focus on data from interviews and a survey about how teachers use AI in their classrooms.

Landscape Highlights

What We Know

- There is much excitement about the potential for AI to improve teaching and learning *and* the growing concern over AI's potential for harm.
- Educators need guidance to evaluate whether AI-generated outputs are accurate, aligned with instructional goals, and appropriate for classroom use.

What Is New

- A structured approach to evaluate AI tools in K–12 classrooms, with an emphasis on transparency in tool development and evaluation.

What is the current landscape of AI tools for K–12 teachers?

The current landscape of AI tools in education includes two broad categories: widely available general-purpose chatbots (e.g., ChatGPT, Microsoft Copilot) and emerging education-specific tools (e.g., MagicSchool, Khanmigo). Although both categories are increasingly visible in classrooms, they differ substantially in how well they align with instructional goals, standards, and classroom realities.

Schools have rapidly adopted general-purpose chatbots mainly because of their free (or cheap) access and ease of use. For example, a nationwide survey found that more than half of U.S. teachers had tried ChatGPT within a few months of its release (Impact Research & Walton Family Foundation, 2023). Teachers have reported using these general AI tools to streamline everyday tasks by prompting chatbots to brainstorm lesson ideas, generate quiz questions or example essays, assist with grading feedback, and even adapt reading materials to different skill levels (Kwid et al., 2024; Uğraş et al., 2024). Systematic reviews have noted that educators see potential uses for chatbots in lesson planning, content creation, and interactive tutoring across subjects, including early literacy and English as a foreign language (Kwid et al., 2024; Zhang & Tur, 2024). Research suggests growing optimism about using AI for elementary reading and writing activities, coupled with careful teacher oversight and parent awareness (Han et al., 2024).

Despite this flexibility and early enthusiasm, general-purpose chatbots are often noted to lack the instructional context needed for effective classroom use. As a result, these chatbots may generate responses that are technically correct but pedagogically misaligned with specific learning objectives, grade levels, or state standards (Jauhiainen & Garagorry Guerra, 2024; Zhang & Tur, 2024). Educators and researchers have flagged issues such as factual errors, unexplained answers, or content that is misaligned with state standards.

In response, a rapidly expanding set of education-specific AI tools has emerged, frequently promising to address these gaps across a wide range of instructional use cases (e.g., lesson planning, assessment, instructional feedback, classroom management). These tools are evolving quickly; new products and features have entered the market and existing tools have frequently expanded their capabilities. This rapid pace of change makes it difficult for educators and researchers to track what tools exist, what they are designed to do, and whether they meaningfully deliver on their claims.

How did we conduct this scan?

To establish a clearer picture of the current landscape, AIR conducted a rapid, light-touch mapping review of teacher-facing AI tools. This approach was informed by scoping and mapping review methodologies, which are designed to systematically identify, categorize, and describe bodies of evidence or artifacts across broad topic areas (Khalil et al., 2025; Campbell et al., 2023). Consistent with rapid review principles, we applied streamlined procedures to enable timely synthesis while maintaining transparency (Wollscheid & Tripney, 2021).

Each tool was independently reviewed by three experts using a coding framework that captured key characteristics, including primary and secondary instructional use cases (e.g., lesson planning, assessment, instructional feedback, classroom management), expected level of teacher engagement, customization options, and the presence of transparency features. Coding was based on information publicly available from tool websites, documentation, and other primary sources describing the tools at the time. Researchers reconciled coding differences through discussion, and results were summarized using frequency counts across tools.

Within the scan, researchers found substantial variation in use cases, including how tools described and labeled their functionalities. To make sense of this variation, tools were grouped according to their primary instructional use cases. Across the landscape, we identified four primary categories:

Instructional planning and content creation: Tools that support lesson planning, material generation, and the creation of instructional resources such as activities, assessments, and aligned lesson structures

Assessment, feedback, and grading: Tools that evaluate student work, generate feedback, or assist in grading across a range of response types

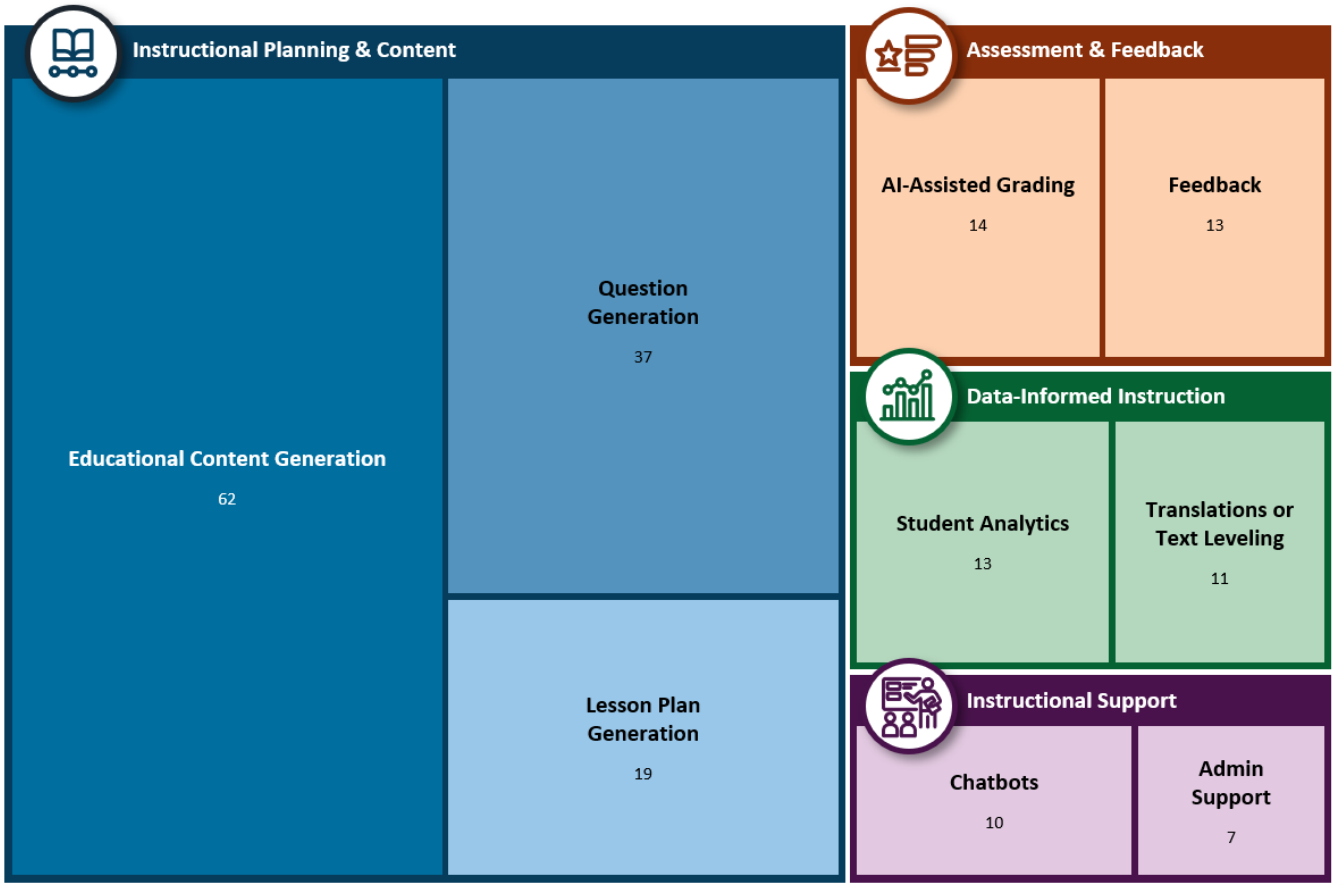
Data-informed instruction and differentiation: Tools that analyze student data or adapt content to support differentiated instruction, including translation and text leveling

Instructional support and classroom operations: Tools that assist with administrative tasks, classroom workflows, or general teaching support, including multipurpose chatbot interfaces

An additional category of tools that focused on integrity and risk mitigation (e.g., AI detection) was identified, but these tools appear less frequently across the current landscape.

A key finding from the scan is that most tools are multifunctional, typically combining two or more use cases within a single platform, and several tools offering four or more use cases. This multifunctionality may increase convenience for teachers, but it also complicates adoption and evaluation. Tools can perform unevenly across different use cases, making it difficult to assess their overall instructional value.

Exhibit 1. Mapping the Use Cases of Teacher-Facing AI Tools



What are key lessons from the landscape review?

After analyzing this collection of teacher-facing AI tools, several lessons emerged:

- **All-in-one tools are convenient but complex.** Most teacher-facing AI tools are offered as suites with multiple features bundled together. While convenient, this makes it difficult to isolate and evaluate individual features or pedagogical benefits.
- **There is little transparency about design and impact.** Information on many tool websites is opaque. Most do not clearly explain their solution approach, the scope of their capabilities, or the outcomes they aim to achieve. In many cases, users must sign up before accessing detailed information, creating barriers for comparison and evaluation.
- **There is little independent evidence.** Across the scan, we did not identify independent evidence supporting the efficacy of most tools. Marketing materials rely heavily on anecdotes and testimonials rather than evidence.

- **There are challenges in evaluating multifunctional tools.** Because tools often serve multiple purposes, evaluating performance is inherently complicated. A tool may perform well in one function and poorly in another.

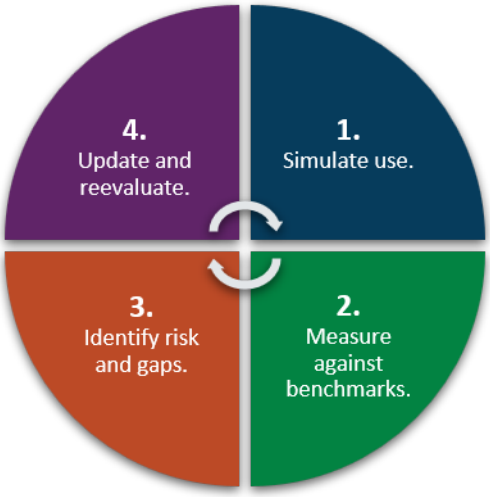
In summary, the scan reveals a rapidly evolving marketplace that is full of promise but lacks clarity and validation. Teachers and district leaders face an overwhelming number of options, while robust guidance on selection and use lags behind.

How will this inform benchmarks and testbed development?

A **virtual testbed** provides a continuous evaluation process in which AI tools can be tested, prompts refined, and outputs assessed. Rather than evaluating a tool one time and moving on, the testbed creates an ongoing process for benchmarking performance, identifying risks, and updating results as tools evolve. The issues identified, including multifunctional complexity, low transparency, and lack of evidence, reiterate exactly *why* this project is focusing on creating education-specific benchmarks and a virtual testbed for AI tools.

This review is the first step in a broader effort to develop expert-informed benchmarks and a continuous evaluation approach that creates consistent, meaningful comparisons across tools.

Exhibit 2. Conceptual Model of the Virtual Testbed



This scan highlights the promise and the limits of current AI tools for teachers.
Although educators have access to an expanding set of AI-powered tools, multifunctional designs, low transparency, and limited independent evidence complicate efforts to understand what works and what does not. These findings reinforce the importance of developing clearer benchmarks and evaluation environments that can support more consistent and meaningful comparisons as AI tools continue to change.

Suggested Citation for This Work

Circi, R., Abeysinghe, B., Sharick, S., Ogut, B. (2026). What's out there? Key insights from a scan of AI tools for K–12 teachers. American Institutes for Research.

References

Campbell, F., Tricco, A. C., Munn, Z., Pollock, D., Saran, A., Sutton, A., White, H., & Khalil, H. (2023). Mapping reviews, scoping reviews, and evidence and gap maps: The same but different? *Systematic Reviews, 12*(1). https://doi.org/10.1186/s13643-023-02178-5

Han, A., Zhou, X., Cai, Z., Han, S., Ko, R., Corrigan, S., & Peppler, K. A. (2024). Teachers, parents, and students' perspectives on integrating generative AI into elementary literacy education. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). ACM. https://doi.org/10.1145/3613904.3642438

Impact Research & Walton Family Foundation. (2023, March 1). *ChatGPT used by teachers more than students* [Survey report]. Walton Family Foundation. https://www.waltonfamilyfoundation.org/chatgpt-used-by-teachers-more-than-students-new-survey-from-walton-family-foundation-finds

Jauhiainen, J. S., & Garagorry Guerra, A. (2024). Generative AI and education: Dynamic personalization of pupils' school learning material with ChatGPT. *Frontiers in Education*, 9**,** 1288723. https://doi.org/10.3389/feduc.2024.1288723

Khalil, H., Welch, V., Grainger, M., & Campbell, F. (2025). Methodology for mapping reviews, evidence maps, and gap maps. *Research Synthesis Methods*, *6*(5), 786–796. doi:10.1017/rsm.2025.25

Kwid, G., Sarty, N., & Yang, D. (2024). A review of AI tools: Definitions, functions, and applications for K–12 education. *AI, Computer Science and Robotics Technology*, *3*(1), 1–22. https://doi.org/10.5772/acrt.20240048

Uğraş, H., Uğraş, M., Papadakis, S., & Kalogiannakis, M. (2024). ChatGPT-supported education in primary schools: The potential of ChatGPT for sustainable practices. *Sustainability, 16***,** 9855. https://doi.org/10.3390/su16229855

Wollscheid, S., & Tripney, J. (2021). Rapid reviews as an emerging approach to evidence synthesis in education. *London Review of Education, 19*(1), 1–17. https://doi.org/10.14324/LRE.19.1.32

Zhang, P., & Tur, G. (2024). A systematic review of ChatGPT use in K-12 education. *European Journal of Education, 59,* e12599. https://doi.org/10.1111/ejed.12599

1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | AIR.ORG

Established in 1946, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit institution that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit AIR.ORG.