AI Hub for Education

The Evidence Base on AI in K-12: A 2026 Review

The existing research on the impacts of AI on students and teachers

Updated November 2026 Includes 32 additional studies



Seth Walker Lily Fesler JP Martinez Claeys Chris Agnew Susanna Loeb

Stanford | SCALE Initiative *Accelerator for Learning*

Executive Summary



The causal evidence on generative artificial intelligence (AI) in education has grown substantially since our March 2026 Evidence Base Review. This update examines new randomized controlled trials and quasi-experimental studies in the AI Hub for Education's Research Repository that were published **between October 2025 and June 2026**. It assesses whether the new evidence supports, challenges, or extends the findings we published in March.

Evidence on how AI impacts K–12 students and educators is growing but remains sparse.

The causal evidence base on AI in education has grown substantially, from 20 papers in the March 2026 Evidence Base Review to 52 papers in this update (an increase of 32 papers). The evidence base has expanded most rapidly for student-facing AI (relative to educator-facing AI), and U.S.-based research has also increased. Five of the new student-facing studies were conducted in U.S. K–12 settings, compared with none in the March 2026 Evidence Base Review. The findings can provide guidance for education leaders, but should not be treated as broadly generalizable across U.S. K–12 settings.

Research on current AI tools suggests the following key findings:

Students

- **AI appears to be better for learning when students use it for learning support than when they use it for producing work**, although neither type of use consistently improves learning. Using AI to offload academic tasks can reduce student learning. The evidence is more favorable (although still mixed) when students use AI explicitly to study and learn.
- **AI tools can improve students' perceived learning experiences, even when they do not improve learning outcomes.** Students often report more positive learning experiences with AI, but those positive experiences often do not translate to stronger learning outcomes. AI support can also improve student well-being and confidence.
- **How AI support is designed and delivered can shape its effects on student learning, but the evidence is still unclear on which characteristics matter most.** Tool features and implementation

support may matter more than tool type. AI-based educational tools do not always outperform general-purpose chatbots. Interactive formats, adaptive sequencing, and tool use guidance show promise for improving use and effectiveness, though studies of when to provide AI support have produced inconsistent results.

- **AI use and AI-generated content may affect motivation, trust, confidence, and well-being.** Educators' use of AI can negatively affect student motivation and confidence, and engaging in independent work before consulting AI can temper trust in it. Structured AI support can improve some aspects of students' well-being and social confidence, but has not consistently improved emotional health outcomes.

Educators

- **AI tools can save teachers time on tasks such as lesson planning, but they can also have unintended effects on students if not used carefully, like undermining student motivation.** Earlier research found that AI reduced lesson-planning time without lowering output quality. However, one new study found that providing teachers with an AI support tool reduced students' motivation and confidence. The difference in benefit may come down to how educators use AI and for what tasks.
- **AI can increase the quantity and quality of feedback educators give to students, help teachers encourage student assignment completion, and give real-time instructional suggestions that improve educator practice.** Educators given AI-generated drafts or suggestions provided more and higher-quality feedback without spending more time on the task. Students who receive AI-generated feedback can perform as well as or better than students who receive educator feedback.
- **Educators' experience and prior performance can influence the effectiveness of AI support.** Earlier research found that targeted, real-time guidance particularly benefited lower-rated and less experienced tutors, but a new study found that a broader teaching-support chatbot reduced achievement among students of lower-performing teachers. Differences in the tools' outputs and the expertise required to evaluate and adapt their support may help explain the divergent findings.

Overall, the updated evidence reinforces the conclusion that the effects of generative AI depend on how it is designed and used. AI shows greater promise when it supports learning and targeted educator tasks than when it replaces student thinking or broader instructional judgment. Tool features and implementation affect how people use AI and how effective it is. More K–12 research is needed to determine which approaches produce lasting benefits, for whom, and under what conditions.

Table of Contents

I. **Introduction** 5

II. **Characteristics of Causal Research on AI in K–12** 8

III. **Key Findings** 13

Students 13

Educators 20

IV. **Conclusion** 24

Appendix 26

A. Summary Table of New Research 26

B. Report Methodology 29

C. Supplemental Research Resources 31

Glossary 32

References 34

Acknowledgments 39

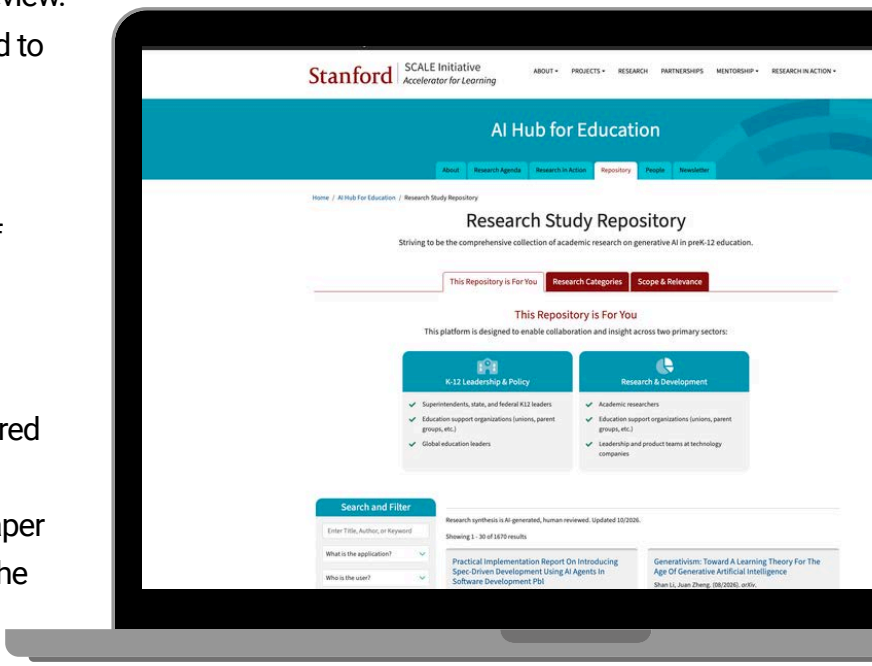
Recommended citation:

Walker, S., Fesler, L., Martinez, J., Agnew, C., & Loeb, S. "The Evidence Base on AI in K–12: A 2026 Review (Update)," AI Hub for Education of the SCALE Initiative, Stanford University, 2026.

Introduction

Education leaders make decisions about whether and how to incorporate generative artificial intelligence (AI), often without easy access to evidence about its effects on students and teachers. To make this evidence easier to find, the AI Hub for Education's Research Repository ("the Research Repository") brings together the research on AI relevant to K–12. The AI Hub published The Evidence Base on AI in K–12: A 2026 Review in March 2026 ("the March 2026 Evidence Base Review"), to complement the Research Repository with an overview of the current causal research and a synthesis of key findings. The report identified 20 strong causal papers in the Research Repository that were published through September 2025.[1] It found mixed outcomes for AI use by students but greater promise for educator use of AI.

The evidence base has expanded substantially since the March 2026 Evidence Base Review. At the end of June 2026, the Research Repository contained 1,686 papers, over twice the 818 papers available in the March 2026 Evidence Base Review.[2] For this update, we examined all papers added to the Research Repository that were published between October 2025 and June 2026. We identified 32 additional strong causal papers, including 28 papers that examined impacts of student-facing AI tools and four papers that examined impacts of educator-facing tools. Notably, five of the new student-facing papers were conducted in U.S. K–12 settings, compared with none in the March 2026 Evidence Base Review, and one additional educator-facing paper was conducted in a U.S. K–12 setting. Given the speed of AI development, we are releasing this update to the 2026 review in advance of the next comprehensive spring 2027 review.



[1] The March 2026 Evidence Base Review included papers published through September 2025 and added to the Research Repository through October 2025, due to a one-month indexing lag.

[2] The total of 1,686 papers includes 1,578 papers in the June 2026 Research Repository update and 108 additional papers identified and added during preparation of the November 2026 Evidence Base Review Update.

Box 1: Methodology & Considerations

How did we identify papers to analyze in this report?

We considered papers in the Research Repository that were published between October 1, 2025 (the cutoff for the March 2026 Evidence Base Review) and June 30, 2026.[3] The Research Repository collects descriptive studies, causal studies, and systematic reviews on the use of AI tools that are relevant for education. It draws primarily from arXiv, as well as papers identified by researchers, practitioners, or SCALE team members.

What do we consider to be "AI" in the Evidence Base reports?

This update focuses on generative AI: tools that generate new content in response to user input, including general-purpose AI tools, AI tutors, and AI-generated feedback systems. The March 2026 Evidence Base Review also focused primarily on generative AI but also included some non-generative tools that we do not include in this update, such as non-generative natural language processing methods to analyze transcripts, score essays, and provide classroom analytics.

What types of studies do the key findings come from?

While other research designs can reveal needs and explain implementation, the report's key findings draw on causal studies of educational interventions, including randomized controlled trials (RCTs) and quasi-experimental designs (QEDs), because these methods provide the strongest basis for estimating impact (See Appendix C.1).

How did we review for paper quality?

We used a two-step review process. First, a large language model ("LLM") preliminarily screened papers for strong causal evidence. Stanford researchers then reviewed prioritized papers for relevance and causal rigor using a structured codebook (See Appendices B and C). Both the classification prompt and codebook are based on the What Works Clearinghouse Procedures and Standards Handbook, Version 5.0 (2022).

How many papers did we identify?

Of the 868 papers added to the Research Repository during the review period, 201 were RCTs or QEDs and 97 advanced to human review.[4] Thirty-two met our standard for strong causal evidence (Appendix C) and were included in the synthesis presented in this report.

---

[3] To incorporate relevant emerging evidence, both reviews included a small number of causal papers identified by the research team that were published after the applicable cutoff dates.

[4] A small number of papers suggested by SCALE researchers during preparation of the November 2026 Evidence Base Review Update did not undergo preliminary LLM screening and proceeded directly to human review.

Box 2: Report Limitations

What are the limitations of this report?

There are four limitations with this report:

- The Research Repository consists primarily of preprints from arXiv, which have not undergone peer review. We include preprints because the publication process can be slow, and limiting the Research Repository to published papers would exclude important recent research. To strengthen confidence in our findings, at least two human researchers review each of the causal papers discussed in the key findings.
- Because our primary screening process focuses on arXiv preprints, we may miss relevant papers published in journals or disseminated through other sources. We may also miss papers because of limitations in our search terms or classification procedures. To mitigate these limitations, we supplement the Research Repository screening process by prompting multiple LLMs to identify additional causal papers of AI in K–12 education.
- Our screening process relies in part on automated classification to prioritize papers for human review. Although human researchers review the papers prioritized for inclusion, the preliminary LLM classification may incorrectly screen out a relevant causal study before it reaches human review.
- Causal evidence does not necessarily imply that findings generalize broadly across educational settings. The papers included in this review vary in their populations, educational settings, AI tools, implementation conditions, and outcomes. Findings should therefore be interpreted in light of the contexts in which individual papers occurred rather than as evidence that the same effects will replicate across all educational settings.

Did we miss a study?

Submit additional or forthcoming research to the Research Repository: scale.stanford.edu/genai/repository/add

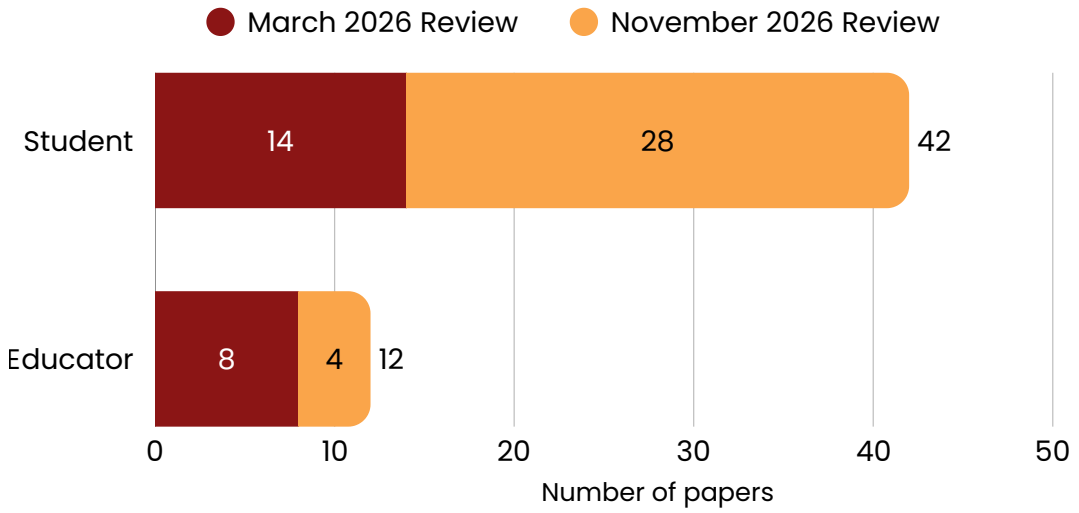
Characteristics of Causal Research on AI in K–12 *(to date)*

The generative AI in education causal evidence base has grown substantially since the March 2026 Evidence Base Review. We identified 32 additional causal papers, bringing the total evidence base to 52 papers.

Most of the newly identified papers examine student-facing generative AI tools.

Of the 32 new high-quality causal papers, 28 examine the effects of student-facing AI tools and four examine the effects of educator-facing tools *(Figure 1)*. Combined with the papers from the March 2026 Evidence Base Review, 42 papers investigate AI use for students and 12 papers investigate educator use.

Figure 1: Number of papers in the Evidence Base Review by the **user of the AI tool**

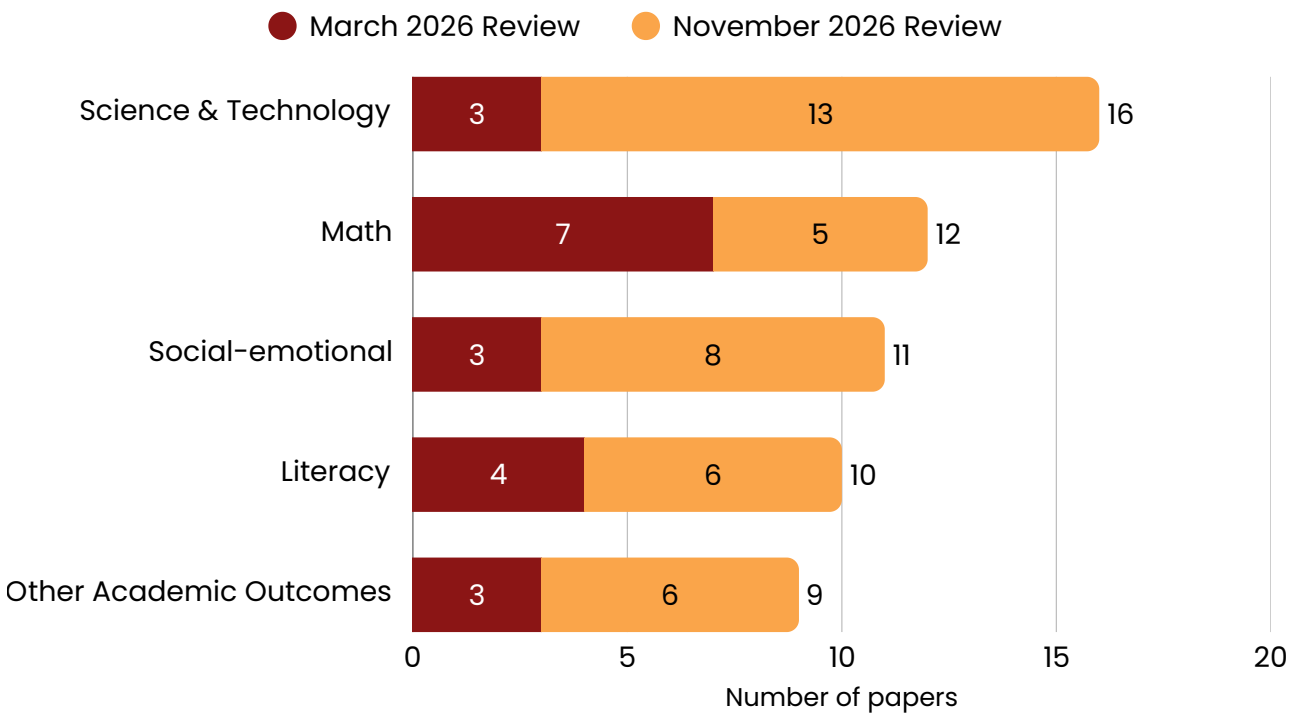


Note: Papers examining tools used by both students and educators are classified in both categories. As a result, the sum of papers across user categories exceeds the total number of unique papers in the evidence base.

The new evidence on academic outcomes primarily focused on science and computer science rather than math or literacy.

Papers were classified based on the outcome studied: literacy, math, science and technology, other academic outcomes, or social-emotional outcomes.[5] "Science & Technology" was the most common outcome category in the update, growing from 3 to 16 papers *(Figure 2)*.[6] This focus may reflect the larger number of papers conducted outside K–12 settings, where papers more often examine science and computer science. Research on social-emotional outcomes also grew substantially, from 3 to 11 papers.

Figure 2: Number of papers in the Evidence Base Review by the **study outcome**



Note: Papers may examine effects on multiple types of outcomes and be included multiple times in this figure. As a result, the sum of papers across study outcomes exceeds the total number of unique papers in the evidence base.

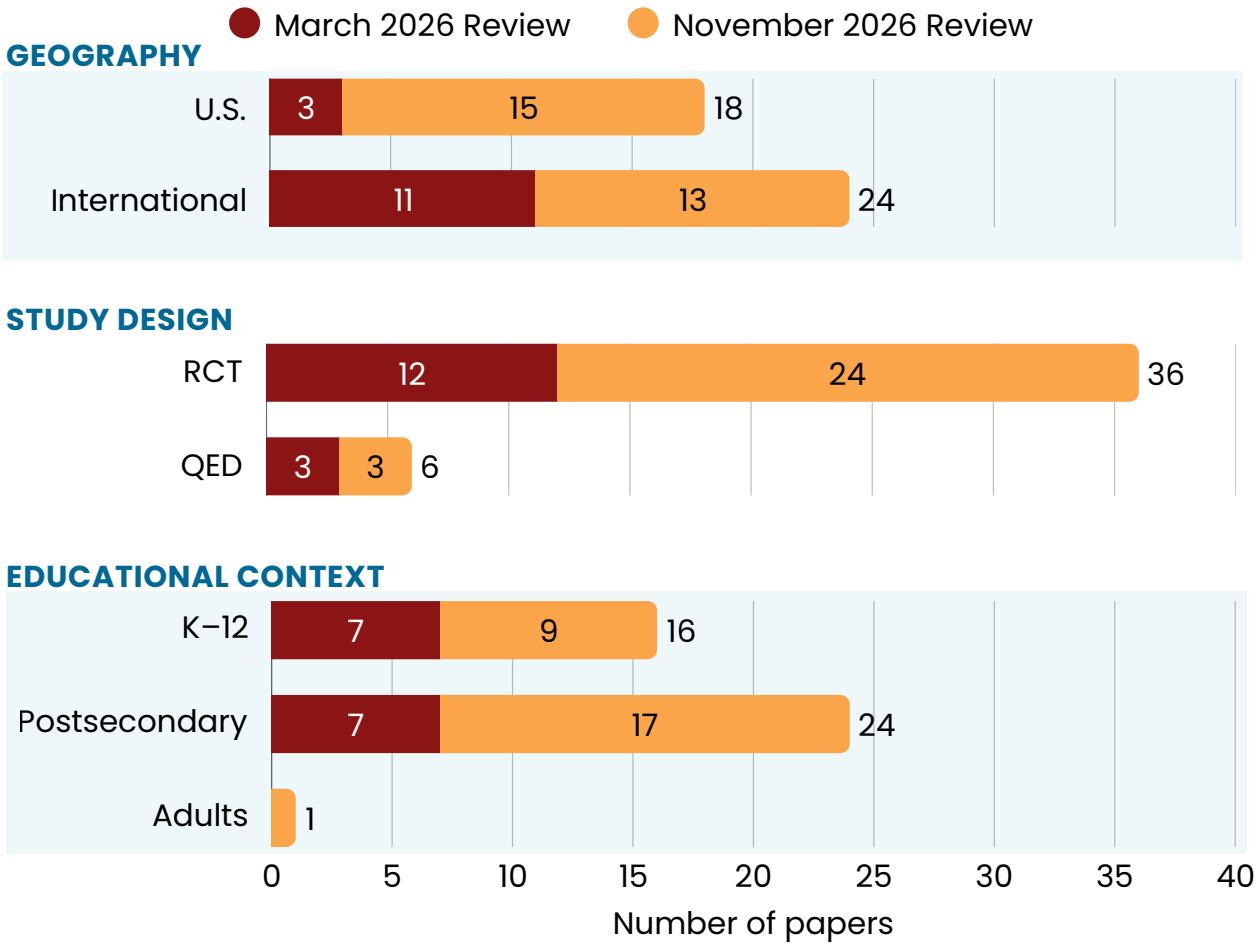
[5] "Other academic outcomes" includes academic subjects not captured by literacy, math, or science & technology, such as social science, history, economics, and design.

[6] Seven of the 13 new papers on science and technology examine computer science or programming outcomes, and six examine natural science outcomes (such as physics).

U.S.-based research on student use of AI has grown considerably, but most of the evidence comes from postsecondary settings.

In contrast with the March 2026 Evidence Base Review, the papers added in the November 2026 Evidence Base Review Update include slightly more papers conducted in the U.S. (15) compared with international settings (13), expanding the U.S. evidence base on students from 3 to 18 papers *(Figure 3a)*. The vast majority of papers continue to use an RCT design, rather than a QED design, to estimate causal effects. Most new papers are in postsecondary settings, as opposed to K–12 settings.

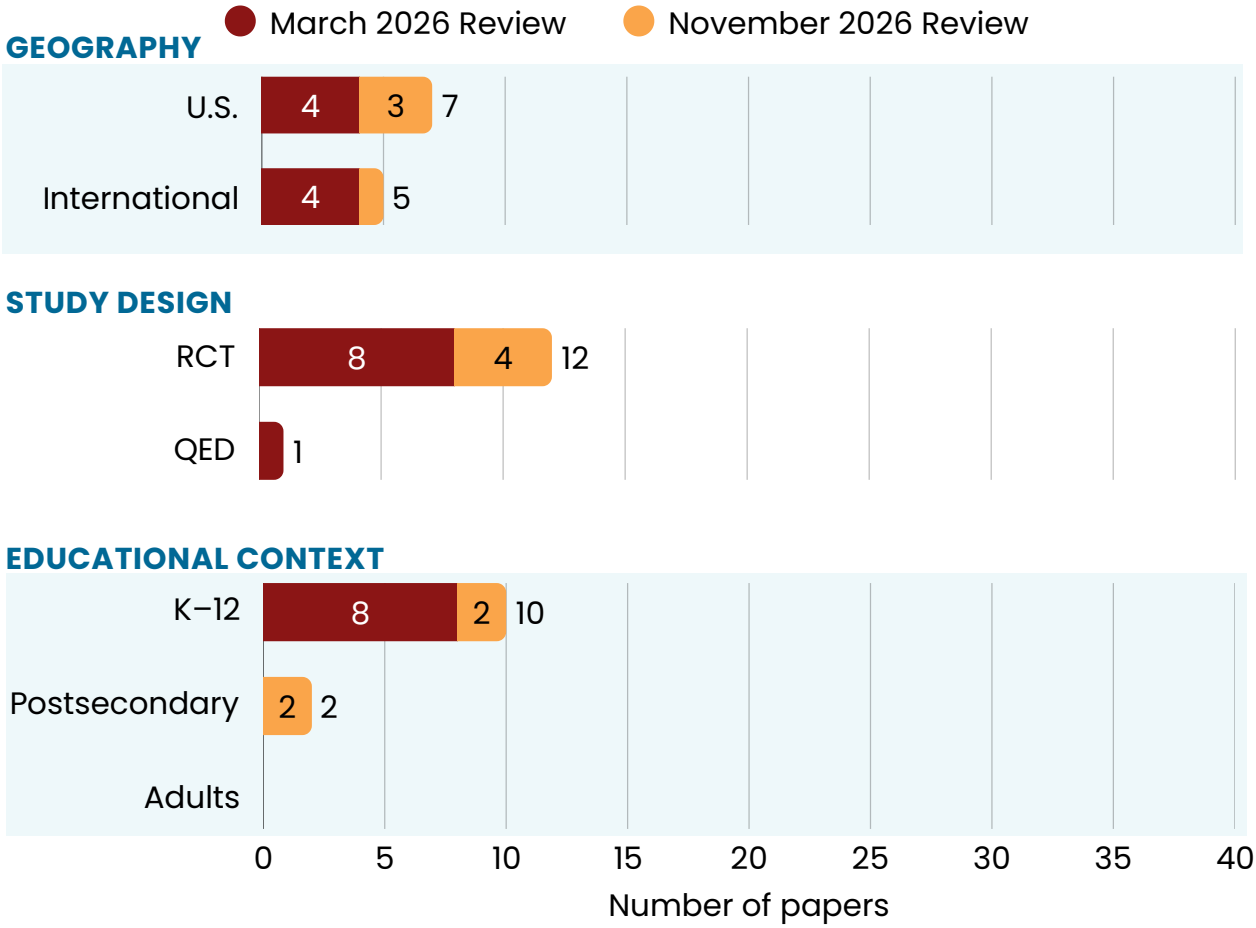
Figure 3a: Number of papers in the Evidence Base Review focused on **student-facing AI tools, by geography, study design, and educational context**



Much of the new educator-facing research is in U.S. and K–12 settings.

Three of the four new papers were conducted in the U.S., and two of the four studied educators teaching or supporting K–12 students. All four new papers use an RCT design to estimate a causal effect of an intervention.

Figure 3b: Number of papers in the Evidence Base Review focused on **educator-facing AI tools, by geography, study design, and educational context**



Overall, the causal evidence base has expanded rapidly, particularly for student-facing tools. At the same time, important gaps identified in the March 2026 Evidence Base Review remain: few causal papers examine educator-facing tools and much of the evidence on student tools comes from postsecondary, rather than K–12, contexts.

Addressing the gaps identified in the March 2026 Evidence Base Review

The March 2026 Evidence Base Review identified important gaps in the causal evidence, particularly around equity and students' social-emotional development. The new evidence begins to address some of these questions, but substantial gaps remain.

The update provides little new evidence on whether AI reduces or reinforces educational inequities.

More research is still needed to understand which students benefit most from different types of AI tools and implementation approaches, how benefits are distributed, and how access to higher-quality tools varies across schools and students.

New papers show some positive effects on social-emotional outcomes for AI tools designed to support well-being, but offer little evidence on how routine use of AI for learning affects students' social-emotional outcomes.

Findings in postsecondary settings suggest that AI tools can have some benefits for resilience, social connection, and social self-efficacy, but have not shown effects for tested measures of depression, anxiety, and stress. Other research provides emerging evidence that educators' AI use can reduce student motivation and that how students use AI may shape their trust in it.

The evidence base still does not show whether AI tools can produce long-term learning gains.

A few papers examine outcomes over a period that extends beyond one-time or short-term interventions, but the evidence does not yet show whether learning gains persist over longer periods.

Key Findings



Students

Students engage with generative AI in different ways. They can use AI to help them complete an assignment, learn new material, receive feedback on their work, or seek social-emotional support. They can also use a variety of different types of tools, including general-purpose chatbots, AI tutoring tools, multimodal AI tools, and AI "well-being" apps. Both how students use AI and how the tools are designed may shape their effects on student learning and other outcomes.

This section reviews evidence on student-facing AI tools, which includes 28 new studies. Together with the studies included in the March 2026 Evidence Base Review, the evidence base now includes 42 student studies. This section also draws on one study of an educator-facing AI tool because it provides new evidence about how teachers' AI use can affect students. We discuss this study more fully in the educator section.

The new research includes five studies conducted in U.S. K–12 settings, compared with none in the March 2026 Evidence Base Review. Although this represents meaningful growth, substantially more research is needed before drawing broad conclusions about AI use in U.S. K–12 schools. Many of the new studies also focus on science and computer science outcomes. The findings below should be interpreted in light of the populations, subjects, and conditions studied.

Number of causal papers that examine student-facing AI tools

14 March 2026 Report
Papers through September 2025

→

42 **November 2026 Update**
Papers through June 2026

Legend

New evidence **supporting** a March 2026 Key Finding

New evidence **conflicting** with a March 2026 Key Finding

New evidence with a **new idea** not covered in March 2026 Key Findings

AI Tools for Producing Work and Student Learning

1 AI appears to be better for learning when students use it for learning support than when they use it for producing work, although neither type of use consistently improves learning.

PREVIOUS FINDING (MARCH 2026 REVIEW) Some student-facing AI tools improve performance while students have access, but effects when students no longer have tool access are mixed.

Definition of Terms

- **AI for learning support** – Students use AI for explanations, hints, practice, or feedback as part of their learning activities.
- **AI for producing work** – Students use AI to generate all or part of an answer or product, replacing work they would otherwise do themselves.

AI for producing work

Supporting

Using AI to outsource academic work can reduce students' performance when they later work without AI.

One study found that AI helped Brazilian undergraduates complete a presentation more quickly but reduced their performance on a surprise conceptual assessment, while a study involving Canadian undergraduates found that beginners who used AI to generate complete coding solutions performed well on the coding task but showed limited understanding of the solution and underlying concepts (Barcaui, 2025; Chen et al., 2026). Another analyzed an adaptive math platform used in U.S. K–12 and postsecondary settings and found that after ChatGPT's release, students spent less time on text-based practice problems that could be easily entered into an AI chatbot than on graph-based problems that were harder to enter (Rismanchian et al., 2026). The same study found that on proctored assessments, where AI use would have been more difficult, students also performed worse on text-based problems relative to graph-based problems. These findings suggest that students may have used AI to bypass some of the problem-solving on text-based practice problems, limiting what they learned from that practice and leaving them less prepared to solve similar problems without AI. An exploratory analysis similarly found that U.S. college students who primarily used AI to draft essays produced higher-quality essays while they had access to AI, but those gains disappeared on a later unaided essay (Contractor & Reyes, 2026).

AI for learning support

Supporting

AI tools designed to explain material, guide practice, or provide feedback in ways that support students' own work show mixed results: some improve learning, while others provide no added benefit over existing approaches.

Immediate learning gains have been observed in K–12 and postsecondary settings across a variety of AI tools designed to support learning, including general-purpose LLMs for self-directed study, Socratic tutors, multimodal study platforms, adaptive practice algorithms, personalized learning support messages, and AI-powered instructional assistants (Anand & Idan, 2026; Chung et al., 2026; Contractor & Reyes, 2026; Heldreth et al., 2026; Kao et al., 2025; LearnLM Team, Google & Fab AI, 2026; Ozturk et al., 2026; Schechter et al., 2025a, 2025b). However, a number of studies have also found no immediate learning benefit from AI tools relative to traditional static materials or standard online practice platforms. These studies span K–12 and postsecondary interventions evaluating real-time AI tutoring tools during academic tasks, on-demand hint assistants, and multimodal learning interfaces (Henze et al., 2026; Rotter et al., 2026; Singh et al., 2026; Taneja et al., 2026). Some studies in the U.S. found that learning gains persisted on follow-up assessments administered three days to one week later (Contractor & Reyes, 2026; Heldreth et al., 2026). These findings provide evidence of short-term retention, but do not establish whether gains persist over longer periods. Another study conducted in K–12 in the U.S. found no statistically significant overall improvement on a delayed learning assessment (Oreopoulos et al., 2026).

General-purpose AI for learning support vs. producing work

Conflicting

General-purpose AI tools have the potential to improve student learning when students use them for learning support instead of producing work.

New studies found that U.S. college students who used general-purpose chatbots to learn unfamiliar material performed better on assessments without access to AI (Anand & Idan, 2026; Contractor & Reyes, 2026). This is in contrast to the studies included in the previous review, which found that Turkish high schoolers and U.S. undergraduates who used general-purpose AI tools performed worse on assessments without AI (Bastani et al., 2025; Kosmyna et al., 2025). One potential explanation for these divergent findings is how students use AI. Students may use it to explain concepts and deepen their understanding or to generate answers and complete assignments. The

exploratory analysis in Contractor and Reyes (2026) is consistent with this interpretation, but does not establish that how students used AI caused the differences in learning. Students' ability to evaluate AI outputs may also matter: one study found that students learned the most from an AI tool if they had more skills with prompting, filtering, and verifying machine-generated outputs (Anand & Idan, 2026). Differences in tools, tasks, student populations, and implementation may also help explain the varying findings across studies.

Student Experiences and Learning

2

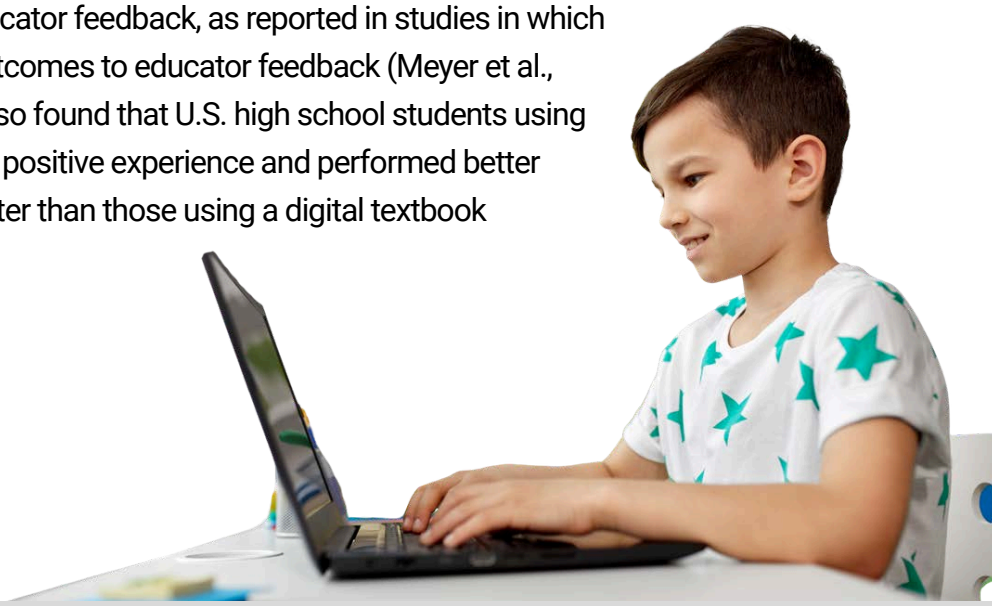
AI tools can improve students' perceived learning experiences, even when they do not improve learning outcomes.

PREVIOUS FINDING (MARCH 2026 REVIEW) AI tools can foster positive experiences in learning, although they may not encourage deeper thinking.

Supporting

Studies in the U.S. and Germany found benefits in undergraduate students' perceptions or workload when AI supported data analysis, conversational study, and programming, but no corresponding improvement in learning outcomes (Gardella et al., 2026; Henze et al., 2026; Taneja et al., 2026).

Both high school and college students have also reported AI feedback to be more useful, motivating, and clear than educator feedback, as reported in studies in which AI feedback led to similar learning outcomes to educator feedback (Meyer et al., 2026; Zhao et al., 2026). One study also found that U.S. high school students using an interactive AI tool reported a more positive experience and performed better both immediately and several days later than those using a digital textbook (Heldreth et al., 2026).



Variation in Design and Delivery of AI Support

3

How AI support is designed and delivered can shape its effects on student learning, but the evidence is still unclear on which characteristics matter most.

PREVIOUS FINDING (MARCH 2026 REVIEW) AI tools can be more effective than general-purpose AI chatbots, and the pedagogical design of the AI tools matters.

Interactive formats, adaptive sequencing, and guidance

New idea

AI tools that incorporate interactive formats, adaptive practice, and structured guidance may improve learning.

A multimodal AI tool outperformed a text-only version among U.S. undergraduates, and an interactive AI tool improved recall relative to a static digital textbook among U.S. high school students (Heldreth et al., 2026; Taneja et al., 2026). However, because the platform in Heldreth et al. (2026) combined multimodal content with feedback, quizzes, and study guidance and did not test these features separately, it is unclear which features drove the learning gains. Tailoring the sequence of practice problems using LLM-derived signals from students' behavior improved performance for Taiwanese high school students relative to a fixed sequence (Chung et al., 2026). Similarly, adding an AI tutoring tool to a computer-assisted math platform for U.S. middle school students produced a significant gain on practiced material only when students were required to answer three questions correctly in a row before advancing (Oreopoulos et al., 2026). In two studies among U.S. postsecondary students, a conceptual roadmap and recommended study sequence produced a significant improvement over unguided use in one study, whereas adding execution tracing to AI-supported programming plans did not improve final coding performance in the other (Anand & Idan, 2026; Jain et al., 2026).

Timing of AI support

New idea

Studies of when to provide AI support have produced mixed results: some studies find benefits from delayed support, while others find no difference.

Two studies of AI feedback timing produced different results: delaying feedback did not affect German high school students' writing performance in one study but improved code quality among adult participants in the U.S. in another (Meyer et al., 2026; Morris & Maes, 2026). Two studies involving postsecondary students in the U.S. found no significant learning benefit from requiring students to attempt a task before receiving AI access or from providing support proactively when students appeared to be struggling rather than waiting for them to request it (Rai et al., 2026; Singh et al., 2026). In contrast, an adaptive approach that determined whether and when AI access became available to university students in Spain produced higher post-test reading scores than unrestricted access, but scores did not differ significantly from those of students without AI access (Rotter et al., 2026).

General-purpose AI vs. educational AI tools

Conflicting

The benefits of general-purpose and educational AI tools may depend on the content and guidance they provide, rather than on tool type alone.

A new study found that U.S. college students who received the correct answer and an explanation immediately during practice (similar to how a general-purpose AI chatbot would respond) performed better than students who worked with chatbots that withheld answers and provided encouragement and hints first (Cao et al., 2026). This contrasts with a study in the previous review, in which Turkish high school students using a general-purpose chatbot performed worse on an exam than those who practiced with a hint-based AI tutoring tool (Bastani et al., 2025). Another new study found similar learning outcomes for students in German middle schools from a general-purpose chatbot and an educational AI tool that added motivational and cognitive-strategy prompts (Fütterer et al., 2026). These differences suggest that adding pedagogical structure alone may not improve learning. The tool that outperformed general-purpose AI in Bastani et al. (2025) used teacher-provided guidance and withheld answers. The other tools used AI-generated guidance that eventually revealed answers (Cao et al., 2026) or added prompts for reflection and learning strategies without changing the underlying instructional content (Fütterer et al., 2026). These findings suggest that educational AI tools are not inherently more effective than general-purpose AI, and that the substance and instructional relevance of the added content may also matter.

NEW TOPIC

Motivation, Trust, and Social-Emotional Outcomes

4 AI use and AI-generated content may affect motivation, trust, confidence, and well-being.

Student motivation, confidence, and trust

New idea

Educators' use of AI may negatively affect student motivation and confidence, and requiring students to work independently before consulting AI may temper their trust in it.

Providing secondary teachers in Türkiye with an AI tool reduced students' intrinsic motivation and confidence, even though students were not told that teachers were using AI (Sungu et al., 2026). Relatedly, adult participants receiving otherwise identical feedback showed greater engagement when it was labeled as human rather than AI, but the differences were not statistically significant after adjustment for multiple comparisons (Morris & Maes, 2026). Additionally, requiring U.S. college students to respond independently before consulting an AI resource reduced how students perceived AI's credibility (Islam et al., 2026).

Authors' Note

The new findings leave substantial uncertainty about which students benefit most or are most harmed and whether AI reduces or reinforces educational inequities.

AI for well-being and social-emotional support

New idea

AI tools designed for well-being and social support improved some outcomes, but benefits did not extend to all measures tested.

One six-week study found that a generative AI well-being mobile app improved U.S. college students' resilience and social connection and helped maintain positive affect and broader measures of well-being, but found no statistically significant effects on depression, anxiety, or stress (Cachia et al., 2025). Another two-week study tested AI companions offering different styles of support including affirmation, empathic questioning, or general support. Each type of companion improved university students' social self-efficacy, while effects on positive and negative affect and other social-emotional skills varied across approaches (Chan et al., 2026). Both studies followed college students over short periods, leaving questions about whether the benefits persist and extend to K–12 students. Because the tools were designed specifically for well-being and social support, these findings do not provide evidence about how routine use of AI for learning affects students' social-emotional development.

Educators



Educators can use AI to support different instructional tasks, including lesson planning, providing feedback to students, and monitoring student progress. These tools vary in how they shape educators' work: some generate materials or feedback that educators can review and revise, while others summarize student data or offer instructional recommendations. The effects of educator-facing AI tools may depend on the task being supported, the quality and specificity of the AI's output, and how educators evaluate and adapt that output.

This section presents findings on how AI use affects educators, including four new studies added for this update. Together with the studies included in the March 2026 Evidence Base Review, the evidence base now comprises 12 studies of educator-facing AI tools. This section also draws on three studies of student-facing AI tools because they provide evidence relevant to educator practice, including comparisons of AI-generated and educator feedback and findings on AI-based learning support.

The evidence base for educator-facing AI remains small. These new studies provide supporting, conflicting, and new evidence related to the March 2026 findings. In particular, they add evidence about impacts of educator AI use on student motivation and confidence and revisit how effects can vary across educator characteristics.

Number of causal papers that examine educator-facing AI tools

8 March 2026 Report
Papers through September 2025

→

12 **November 2026 Update**
Papers through June 2026

Legend

New evidence **supporting** a March 2026 Key Finding

New evidence **conflicting** with a March 2026 Key Finding

New evidence with a **new idea** not covered in March 2026 Key Findings

Lesson Preparation

1 AI tools can save teachers time on tasks such as lesson planning, but they can also have unintended effects on students if not used carefully, like undermining student motivation.

PREVIOUS FINDING (MARCH 2026 REVIEW) AI can reduce teacher time spent on routine tasks, or shift teacher effort, with no evidence of quality losses.

New idea

One new study found that giving teachers in Türkiye access to a broad AI teaching-support tool reduced student motivation and confidence, but had no statistically significant effect on test scores overall (Sungu et al., 2026). An earlier study in England found that AI use (primarily to create questions or quizzes) reduced teachers' lesson-preparation time without a detected decline in lesson-plan quality (Roy et al., 2024). These findings underscore the importance of examining student outcomes alongside teacher efficiency and the quality of teaching materials. In the newer study, although the tool could support a wide range of teaching tasks, teachers used it primarily to generate teaching materials that were accepted with minimal revision. The authors interpret this pattern as consistent with delegating instructional functions to AI, with relatively little engagement in reviewing and adapting the materials it generated. This interpretation is analogous to the possible role of producing work in the student studies: in both cases, greater reliance on AI may reduce the human effort and judgment that shape the quality of the outcome. It also suggests that how teachers evaluate and refine AI-generated materials may matter, although the study does not establish that either pattern caused the effects on students.



Feedback and Instructional Support

2 AI can increase the quantity and quality of feedback educators give to students, help teachers encourage student assignment completion, and give real-time instructional suggestions that improve educator practice.

PREVIOUS FINDING (MARCH 2026 REVIEW) AI can deliver automated feedback on teaching and student progress, improving instructional quality and student outcomes. AI can offer real-time instructional suggestions that improve educator practice and student outcomes, particularly in messaging-based settings.

AI for progress monitoring

New idea

AI-generated summaries of student progress for teachers can increase assignment completion, even if they do not improve academic performance.

In one study, K–12 students in the U.S. were more likely to start and complete assignments if their teachers received AI summaries of assignment reports (Lim et al., 2026). The AI summaries included an overview of the assignment score and completion rate, common learning gaps, students who could use additional support, and instructional recommendations (such as concepts to review). However, class-level assignment scores did not improve, even though more students started and completed the assignments.

AI to support educator feedback to students

New idea

AI may enable educators to provide more high-quality feedback.

AI-generated drafts increased the likelihood that teaching assistants at a U.S. university provided feedback to students and led to longer comments (Mahinpei et al., 2026). Students rated this feedback as similarly useful to feedback written without AI assistance. Another study at a university in the U.S. found that teaching assistants who received AI suggestions gave nearly twice as much feedback, with higher-quality feedback and broader coverage of rubric criteria, without increasing the time spent grading submissions and providing feedback (Lu et al., 2026). Students receiving this AI-assisted feedback produced higher-scoring revisions, but did not perform significantly better on a subsequent post-test.

AI for direct student feedback

Supporting

AI may support educators by generating student-facing feedback that produces outcomes similar to educator feedback. More personalized AI support may also modestly improve academic outcomes.

Real-time multimodal AI feedback produced learning gains similar to prewritten educator comments and was rated clearer, more specific, and more motivating by college learners in the U.S. (Zhao et al., 2026). Results were similar in German secondary schools, where AI-generated and teacher-provided writing feedback produced no statistically significant differences in revisions or subsequent writing performance (Meyer et al., 2026). Beyond feedback alone, individualized AI messages combining performance feedback, motivational prompts, and study guidance modestly improved midterm performance and final-exam participation in a large online course in Türkiye, but did not significantly improve final exam scores (Ozturk et al., 2026).

Educator Characteristics

3

Educators' experience and prior performance can influence the effectiveness of AI support.

PREVIOUS FINDING (MARCH 2026 REVIEW) AI pedagogical supports appear to be most beneficial to lower-rated and less experienced educators.

Conflicting

A new study found that a general teaching-support tool that was primarily used to generate instructional materials had no overall effect on achievement and reduced achievement among secondary students of below-median-performing classroom teachers in Türkiye (Sungu et al., 2026). This is in contrast to a study included in the previous review that provided K–12 tutors in the U.S. with real-time, expert-like instructional suggestions and improved student mastery, with the largest gains for less-experienced and lower-rated tutors (Wang et al., 2025). Differences in the tools, teaching tasks, and the expertise required to use the AI support may help explain the contrasting results. However, the studies do not isolate which factors account for the different results.

Conclusion



The growing causal evidence base identifies uses of generative AI in education that can improve teaching and learning, but substantial uncertainty remains about which approaches are effective and which are detrimental across K–12 settings. The addition of 32 causal papers since the March 2026 Evidence Base Review more than doubled the evidence base and extended it across a wider range of academic and social-emotional outcomes. Much of the new research remains focused on student-facing tools in postsecondary settings. Relatively few papers examine K–12 classrooms or educators' use of AI, and differences across tools, populations, and implementation contexts limit the generalizability of individual findings.

The new evidence suggests that AI is more promising for learning when students use it to explain concepts, guide practice, or provide feedback on their own work than when they use it to generate answers or complete work on their behalf. However, learning gains are not consistent even when AI is used for learning support. Differences in tools, tasks, student populations, and implementation also make it difficult to determine whether the way students use AI explains the varying results. Interactive formats, adaptive sequencing, and guidance on how to use AI show promise, but the evidence is not yet sufficient to identify which design features matter most. Positive student experiences also do not necessarily translate into stronger learning outcomes.

For educators, studies have found that AI can reduce lesson-preparation time without a detected decline in lesson-plan quality, increase the quantity and quality of feedback, and increase student assignment completion. However, one study of a broad teaching-support tool found reduced student motivation and a marginal reduction in confidence. That study found no statistically significant overall effect on test scores, but scores declined among students of lower-performing teachers. These findings suggest that evaluating AI access alone is insufficient; research must also examine the specific tasks AI supports, the guidance it provides, and how students and educators incorporate that support into their work.

For education leaders, these findings support evaluating specific AI tools and uses against their intended goals and in the setting where they will be implemented. For example, a school using AI to help teachers give feedback could assess whether teachers provide more useful feedback in the time available and whether students perform better on subsequent independent work. Such evaluation should consider student learning and motivation alongside teacher workload and the quality of instruction.

Future studies could advance our understanding by prioritizing K–12 settings, measuring learning after AI access is removed, and examining whether effects persist and transfer beyond the immediate task. Further research could clarify when AI reduces or reinforces educational inequities, including how access to effective tools and their benefits and harms vary across students and schools. More causal research is also needed on educator-facing tools, including how teachers' expertise shapes their ability to evaluate and adapt AI-generated support. As the evidence base develops, studies should assess outcomes beyond academic performance, including motivation, trust, confidence, well-being, and the quality of interactions among students and educators. The central question is therefore not simply whether generative AI is effective in education, but when, how, and for whom it supports learning—and when it may undermine it.



Appendix A

Summary Table of New Research

Legend

Randomized Controlled Trial (RCT) | Quasi-Experimental Design (QED) | Peer-Reviewed Research

Paper	User	Intervention Description	Sample	Key Finding
Anand & Idan (2026)	Students	Students studied ML and LLM topics for three days using ChatGPT or non-LLM resources. Novice ChatGPT users received conceptual scaffolding, extra time, peer collaboration, or no added support.	179 predominantly engineering undergraduates at a U.S. university	ChatGPT access improved knowledge scores. Conceptual scaffolding offered marginal benefits for novices, while students with greater AI interaction competence benefited more. GPA and prior knowledge did not moderate gains.
Cao et al. (2026)	Students	Students received GPT-5 feedback on multimedia-principle tasks either immediately with answers and explanations or progressively through encouragement, prompts, and hints before answers.	199 U.S. college students recruited through Prolific	Immediate, non-layered feedback yielded higher post-test scores. Layered feedback increased perceived encouragement, independence, mental effort, and repeated task submissions.
Chan et al. (2026)	Students	Students discussed personal social dilemmas daily for 14 days over text with affirmative, empathic-questioning, or general-support AI companions, while control students journaled without AI.	134 university students	All AI companion groups improved social self-efficacy similarly. Affirmative AI most improved positive affect, while empathic-questioning AI showed the broadest social-emotional gains and greatest reduction in negative affect.
Chen et al. (2026)	Students	Beginner and intermediate programming students completed a coding task using either ChatGPT as their sole external resource or conventional non-AI online resources.	24 Canadian undergraduates in a programming course	ChatGPT improved beginners' performance without corresponding understanding gains. Complete AI-generated solutions often lacked understanding. Intermediate students generally understood more with conventional resources.
Chung et al. (2026)	Students	During a five-month Python course, students used the same GPT-4o tutor but received either adaptive practice sequencing based on learning behaviors or fixed easy-to-hard sequencing.	770 Taiwanese high school students	Adaptive sequencing raised unassisted certification scores by 0.15 SD, especially for beginners and lower-tier schools. Greater engagement, rather than problem quantity or difficulty, may explain gains.
Contractor & Reyes (2026)	Students	Students learned about an unfamiliar technical topic and wrote an analytical essay using either generative AI or traditional resources.	211 U.S. undergraduates	AI improved knowledge scores immediately and one week later. Unaided writing improved in style, clarity, and relevance, but not overall quality. Gains persisted more with explanatory than essay-drafting use.
Fütterer et al. (2026)	Students	Students used a GPT-4o chatbot with utility-value prompts, cognitive-strategy prompts, or standard ChatGPT interactions while studying physics or English.	241 German students in grades 7–9	Cognitive-strategy prompts reduced perceived relevance. Neither prompting approach significantly improved strategy use, interest, effort, or domain knowledge relative to standard ChatGPT.
Gardella et al. (2026)	Students	Students completed programming tasks with GitHub Copilot or a human partner, then repeated tasks individually without assistance approximately one week later.	22 novice and intermediate programmers at a U.S. university	Copilot improved assisted task performance and reduced workload, but not unassisted retest performance. Performance declined more after Copilot, while human collaboration generated more positive emotional experiences.
Heldreth et al. (2026)	Students	Students learned neuroscience using either Learn Your Way, which adapted textbook content and added quizzes, feedback, and review prompts, or a static digital textbook.	58 U.S. high school students recruited through Bold Insight	The AI platform improved immediate and 3–7-day follow-up assessment scores relative to the textbook and produced more positive reported learning experiences.
Henze et al. (2026)	Students	Student teachers analyzed physics data using either ExperiMentor, a GPT-4o-based chatbot, or Excel.	50 student teachers at a German university	Learning gains did not significantly differ, but ExperiMentor users reported greater enjoyment, perceived effectiveness, and preference for the method.

Islam et al. (2026)	Students	Students used AI under one of three workflows: answer independently before consulting AI, verify AI outputs, or use AI without restrictions.	42 U.S. undergraduates in a visual analytics course	Students required to answer independently before consulting AI rated it as less credible than those with unrestricted access. Credibility rating did not significantly differ for the verification group.
Jain et al. (2026)	Students	Students developed coding plans with iterative AI feedback. The treatment group additionally used execution tracing to evaluate their plans before coding. Both groups then coded without AI assistance.	20 undergraduate / graduate students with introductory CS proficiency	Execution tracing shifted plans toward fewer, more goal-driven steps but did not significantly improve programming performance compared with AI-supported planning alone.
LearnLM Team, Google & Fab AI (2026)	Students	For eight weeks, teachers integrated Gemini Guided Learning into half their mathematics lessons. Students worked in pairs using teacher-set objectives, scaffolded prompts, and class discussions, versus standard instruction.	1,423 grade 7–8 mathematics students in Sierra Leone	Guided Learning increased mathematics scores by 0.258 SD. Effects were larger among students with higher baseline mathematics performance.
Meyer et al. (2026)	Students	Students received feedback on English essay drafts from teachers or GPT-4, with AI feedback delivered immediately or after the same delay as teacher feedback.	391 German students in grades 10 and 11	Revision quality and subsequent writing performance did not significantly differ. Immediate AI feedback was perceived as more useful and motivating but did not increase revision time or editing extent.
Morris & Maes (2026)	Students	After a coding tutorial, participants received identical LLM-generated feedback labeled as AI or human and delivered immediately or after a delay.	148 adult Prolific crowdworkers	Delayed versus immediate AI feedback yielded more complex code. Human-labeled feedback did not significantly improve engagement after statistical correction. Doubting the human label was linked to disengagement and less complex code.
Oreopoulos et al. (2026)	Students	Students practiced math on a computer-assisted learning platform, with or without an AI tutor providing hints and explanations. Students were also independently assigned to mastery-based or flexible progression.	6,997 U.S. students in grades 6-8 across 20 schools	AI slowed practice and reduced questions completed without improving overall delayed learning. Under mastery progression, it improved correctness on one practiced item, but not the broader learning measure.
Ozturk et al. (2026)	Students	Online economics students received personalized GPT-4o learning-support messages throughout the semester, including feedback, motivation, study advice, and assessments, versus standard course materials.	28,833 undergraduates in an online economics course in Türkiye	AI-generated messages modestly increased midterm scores and final-exam participation but did not significantly affect final-exam scores or midterm participation.
Rai et al. (2026)	Students	Programming students used the same GPT-4o CodeTutor with Socratic hints, initiated proactively or only in response to student requests.	90 introductory programming students at a U.S. university	Proactively initiated AI support did not significantly improve computational-thinking growth compared with support provided only upon student request, either for conceptual reasoning or code implementation.
Rismanchian et al. (2026)	Students	Researchers analyzed ALEKS math data before and after ChatGPT's release, comparing text-based problems more easily entered into AI chatbots with graph-based problems that were harder to enter, across proctored and unproctored settings.	3.2 million ALEKS interactions from U.S. grade 5–college students	After ChatGPT's release, older students spent less time on AI-susceptible problems without proctoring. On proctored retention items, correct-response odds fell approximately 25% relative to AI-resistant problems.
Rotter et al. (2026)	Students	During three reading tasks, students received chatbot access controlled by reinforcement learning (RL), unrestricted chatbot access, or no chatbot access.	105 undergraduate and graduate students in Spain	RL-timed access improved post-test scores and metacognitive accuracy relative to unrestricted access, but did not significantly outperform no access.
Singh et al. (2026)	Students	Students wrote debugging hints without AI, with immediate on-demand GPT-4 hints, or with AI hints available only after submitting their own hints.	97 U.S. graduate students in a data science course	Deferred AI support improved hint quality and error identification while reducing irrelevant detail. On-demand support also improved hint quality. Analyzed post-test learning outcomes did not significantly differ across conditions.

Taneja et al. (2026)	Students	Participants studied meiosis using textbook-grounded conversational AI with text and images, text-only conversational AI, or non-conversational textbook search.	124 participants with a biology major recruited via Prolific	Students using multimodal AI scored higher than those using text-only AI, but not textbook search. Both conversational tools were rated easier and more engaging despite text-only AI's lowest scores.
Zhao et al. (2026)	Students	Students received GPT-5-generated multimedia feedback, including personalized text, slides, and optional audio, or feedback written by a human educator.	197 U.S. college students recruited through Prolific	AI and educator feedback produced statistically equivalent learning gains. Students rated AI feedback clearer, more specific, more motivating, and less cognitively demanding, with similar trust and perceived correctness.
Barcaui (2025)	Students	Students prepared an AI/ML presentation using unguided ChatGPT or traditional non-AI study resources.	85 Brazilian undergraduate business students	Students using ChatGPT reported substantially less preparation time (3.2 versus 5.8 hours), while students using traditional resources scored higher on a surprise delayed conceptual understanding test 45 days later.
Cachia et al. (2025)	Students	Students received six weeks of Flourish, a generative AI well-being app combining coaching conversations, emotion check-ins, activities, and reflections, versus care-as-usual.	486 undergraduate students across three U.S. institutions	Compared with care as usual, Flourish improved resilience, belonging, and loneliness and preserved some positive well-being measures, but did not significantly improve depression, anxiety, stress, negative affect, or academic self-efficacy.
Kao et al. (2025)	Students	Students completed inquiry activities with ChatGPT Study Mode's Socratic prompts and feedback, peer feedback without AI, or no intervention.	90 10th-grade students at a Taiwanese high school	Students in the AI condition scored higher on argumentation and critical thinking than both comparison groups and reported greater self-efficacy and engagement. Metacognitive self-regulation did not significantly differ across conditions.
Schechter et al. (2025a)	Students	Students used Coursemojo for curriculum-integrated reading and writing, with AI questioning, scaffolding, and feedback, or received standard English language arts instruction.	2,771 grade 6 students across 13 U.S. schools	Coursemojo students scored 10.5 points higher on STAAR reading and language arts and were more likely to pass. NWEA MAP performance did not significantly differ.
Schechter et al. (2025b)	Students	Students used Coursemojo for curriculum-integrated reading and writing, with AI questioning, scaffolding, and feedback, or received standard English language arts instruction.	2,045 grade 6 students across 12 U.S. schools	Coursemojo students scored approximately eight points higher on Tennessee's TCAP ELA assessment and performed better on Aimsweb and TN Ready assessments. More frequent use was associated with higher TCAP scores.
Lim et al. (2026)	Educators	Teachers received standard ASSISTments reports or reports with on-demand AI summaries identifying learning patterns, struggling students, and instructional suggestions over eight weeks.	874 U.S. teachers	Despite no significant effect of AI summaries on teachers' logged behavior, their students started and completed more assignments, though the value-added measure did not significantly improve.
Lu et al. (2026)	Educators	Teaching assistants (TAs) graded essays with or without AI-suggested rubric comments they could adopt, edit, or discard.	11 TAs and 354 U.S. undergraduates in an economics course	AI-assisted TAs provided more and better-rated feedback without taking longer. Students receiving AI-assisted feedback produced stronger revisions but did not score significantly higher on a later post-test.
Mahinpei et al. (2026)	Educators	After grading submissions, TAs received optional AI-drafted feedback comments they could use, edit, or ignore, compared with no draft.	11 TAs and 88 U.S. undergraduates in an ML course	AI drafts increased the likelihood of providing feedback by about 11 percentage points and increased comment length. Time per character and rated usefulness were similar.
Sungu et al. (2026)	Educators	Teachers received a GPT-4o teaching-support chatbot for lesson planning, assessment, feedback, differentiation, and administration for one semester, or continued usual practice.	193 teachers and 2,816 students in grades 5-12 in Türkiye	Teacher AI access reduced student motivation and confidence but not average exam scores. However, scores fell for students of lower-performing teachers.

*To see a full list of all 52 studies included in both the original March 2026 report and the November 2026 update with filter options available and details about each paper, see **here**.*

Appendix B
Report Methodology

This appendix briefly describes the criteria used to determine a paper's eligibility for inclusion in the Evidence Base Review and the process for translating the research into key findings. For a detailed overview of the Research Repository, including how papers are collected and organized, see Appendix B of the March 2026 Evidence Base Review. For the rationale for focusing on randomized controlled trials (RCTs) and quasi-experimental designs (QEDs), see Appendix C.1 of the March 2026 Evidence Base Review.

Randomized Controlled Trials (RCTs) are studies in which participants (such as students, classrooms, or schools) are randomly assigned to receive an intervention or to serve as a comparison group. Because the assignment is random, the two groups are similar on average before the intervention begins, allowing researchers to confidently attribute any differences in outcomes to the intervention itself rather than to pre-existing differences.

Quasi-experimental designs (QEDs) evaluate an intervention without random assignment, using a carefully selected comparison group to estimate its effects. Researchers use design strategies and statistical methods to estimate causal effects as well as possible. These studies can support reasonable conclusions about whether the intervention contributed to observed differences in outcomes, even when randomization is impractical or unethical.

B.1 Determining eligibility of papers into the Evidence Base Review

For the November 2026 Evidence Base Review Update, we followed an improved upon structure to the March 2026 Evidence Base Review. As with the previous report, the review of key findings considers papers assessed as having sufficient methodological rigor to support credible causal conclusions.

To determine eligibility and assess rigor, we followed guidance from the What Works Clearinghouse Procedures and Standards Handbook, Version 5.0 (2022) ("WWC Handbook"). The WWC Handbook provides guidance for evaluating whether education studies provide causal evidence.

We prompt Gemini 3.0 Flash to classify papers based on their methodology. Drawing on Section 3 of the WWC Handbook, this step screens studies for key methodological features relevant to WWC standards, including whether the design supports an appropriate comparison between treatment and control conditions, whether major threats to validity (e.g., confounding) are addressed, and whether the reported outcomes and analyses are suitable for estimating intervention effects.

- We rated papers on a 1–3 scale across the following dimensions: construct validity, outcome reliability, outcome alignment, outcome data collection consistency, confounding factors, assignment to conditions, compositional change, and baseline equivalence. To avoid excluding potentially relevant studies, a study did not pass this step only if it received a 1 on any dimension; 81 of 185 papers passed this stage.

After classification, Stanford researchers evaluated the 81 papers that passed the LLM screening, along with 16 additional papers identified during preparation of the report that had not undergone LLM screening, for a total of 97 papers. Researchers followed the questionnaire included in Appendix C, which is based on the WWC Handbook classifying papers as Include, Exclude, or Needs Additional Review. For every paper that is classified as Needs Additional Review, two additional researchers reviewed the paper to determine inclusion. The key findings in this update are developed from papers classified as Include, either initially or after additional review.

B.2 Translating the evidence from selected papers into key findings

Papers passing the process described above were synthesized into the following structure:

- *Key findings:* Provides a high-level summary of what the new evidence allows us to conclude. It may introduce a new takeaway or add nuance to an existing one.
- *Supporting evidence:* Describes the evidence underlying the takeaway, drawing on the new papers and explaining how their findings relate to the March 2026 Evidence Base Review. The new evidence is classified as:
  - Supporting: New findings support a key finding from the March 2026 Evidence Base Review, and no study provides conflicting evidence.
  - Conflicting: New findings conflict with a key finding from the March 2026 Evidence Base Review. Evidence was considered conflicting when studies found effects in opposing directions to those in the March 2026 Evidence Base Review. Differences in the magnitude of effects in the same direction were not considered conflicting because interventions may vary in their context, duration, implementation, or outcomes measured.
  - New: The newly selected paper(s) provide evidence on a topic not covered in the March 2026 Evidence Base Review.

Each key finding and supporting evidence section was drafted by at least one researcher, verified by a second researcher, and reviewed by at least two other members of the Stanford SCALE research team.

Appendix C
Supplemental Research Resources

The resources below offer greater insight into the studies included in review analyses and exactly how new studies were selected and reviewed.

Resource List

Label	Source File
Full List of Causal Papers (includes all *Evidence Base on AI in K–12* reports)	Spreadsheet file (living)
Research Study Eligibility and Evidence Extraction Questionnaire (November 2026) Researchers used this questionnaire to screen studies for eligibility, document study and sample characteristics, evaluate the credibility of outcome measures and research designs, and extract findings from eligible studies. The version sourced here converts the original digital labeling form dated July 30, 2026, into a report-ready format. Conditional prompts and decision rules are shown alongside the questions they follow.	PDF file (static)

Glossary

Term	Definition
Adaptive sequencing	An approach that changes the order of practice or content using information about learner behavior or performance.
AI-assisted feedback	Feedback that an educator produces with AI-generated drafts or suggestions while retaining responsibility for review and delivery.
AI for learning support	Students use AI for explanations, hints, practice, or feedback as part of their learning activities.
AI for producing work	Students use AI to generate all or part of an answer or product, replacing work they would otherwise do themselves.
AI-generated feedback	Feedback produced by an AI system and provided either directly to students or through an educator.
AI tool	A generative AI system or application used to support a task, interaction, or educational activity.
AI tutoring tool	An AI tool that supports learning through instructional interaction, guided practice, hints, explanations, feedback, or other scaffolding.
Causal study	An empirical study that uses a design intended to estimate the effect of an intervention, including a randomized controlled trial or quasi-experimental design.
Educational AI tool	An AI tool designed for instructional use, often with pedagogical guidance, constraints, prompts, or structured learning activities.

Glossary

Term	Definition
General-purpose AI chatbot	A broadly available conversational AI tool that is not designed exclusively for a specific educational task or instructional role.
Generative AI	AI that generates new content in response to user input, such as text, explanations, feedback, or instructional materials.
Multimodal AI tool	An AI tool that uses or presents more than one mode of information, such as text, audio, images, or other visual content.
Quasi-experimental design (QED)	A causal research design that estimates intervention effects without random assignment by using a comparison strategy intended to approximate an experiment.
Randomized controlled trial (RCT)	A study in which participants or units are randomly assigned to intervention and comparison conditions to estimate causal effects.
Strong causal evidence	Evidence from a study that meets the report's standards for relevance and causal rigor after automated screening and human review.

References

Anand, B., & Idan, L. (2026). Generative AI and the productivity divide: Human-AI complementarities in education [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.18143

Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. Social Sciences & Humanities Open, 12, Article 102287. https://doi.org/10.1016/j.ssaho.2025.102287

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Cachia, J. Y. A., Zhao, X., Hunter, J., Wu, D., Lin, E., & De Freitas, J. (2025). AI for proactive mental health: A multi-institutional, longitudinal, randomized controlled trial [Preprint]. SSRN. https://doi.org/10.2139/ssrn.5718163

Cao, J., Zhao, C. Q., Schunn, C., McLaughlin, E. A., Lin, J., & Koedinger, K. R. (2026). Assessing the impact and underlying pathways of sequenced AI feedback on student learning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2604.07469

Chan, S.-L., Zhao, J., Gao, J., Lyu, X., & Wei, J. (2026). Beyond generic answers and empty praise: The differential impact of AI companionship roles on social-emotional learning in university student intervention. In E. G. Blanchard et al. (Eds.), Artificial intelligence in education: AIED 2026 (LNAI 16584, pp. 500–514). Springer. https://doi.org/10.1007/978-3-032-29763-1_34

Chen, R., Jiang, S., Shen, J., Moon, A., & Wei, L. (2026). Examining the usage of generative AI models in student learning activities for software programming. arXiv. https://doi.org/10.48550/arXiv.2511.13271

Chung, A. T.-H., Zhang, B., Kung, L.-C., Bastani, H., & Bastani, O. (2026). Effective personalized AI tutors via LLM-guided reinforcement learning [Preprint]. SSRN. https://doi.org/10.2139/ssrn.6423358

Contractor, Z., & Reyes, G. (2026). Experimental evidence on the learning impact of generative AI [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.08849

Fesler, L., Martinez, J., Agnew, C., & Loeb, S. (2026). The evidence base on AI in K–12: A 2026 review. AI Hub for Education, SCALE Initiative, Stanford University. https://scale.stanford.edu/sites/default/files/The%20Evidence%20Base%20on%20AI%20in%20K-12%20Report.pdf

Fütterer, T., Bardach, L., Kuhn, J., Keller, S. D., & Gerjets, P. (2026). Enhancing school students' self-regulated learning through generative AI support: A randomized controlled trial. Educational Psychology Review, 38, Article 42. https://doi.org/10.1007/s10648-026-10133-8

Gardella, N., Prather, J., Leinonen, J., Denny, P., Pettit, R., & Riggs, S. L. (2026). Fast and forgettable: A controlled study of novices' performance, learning, workload, and emotion in AI-assisted and human pair programming paradigms [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2604.18538

Heldreth, C., Akrong, D., Vardoulakis, L. M., Miller, N. E., Haramaty, Y., Hackmon, L., Belinsky, L., Ortiz Tapia, A., Tootill, L., & Siebert, S. (2026). An experimental evaluation of an AI-powered interactive learning platform. Frontiers in Artificial Intelligence, 9, Article 1783117. https://doi.org/10.3389/frai.2026.1783117

Henze, J., Lademann, J., Becker-Genschow, S., & Bresges, A. (2026). AI-supported data analysis boosts student motivation and reduces stress in physics education. Frontiers in Education, 11, Article 1719670. https://doi.org/10.3389/feduc.2026.1719670

Islam, M. T., Akgun, M., & Billah, S. (2026). Shaping credibility judgments in human-GenAI partnership via weaker LLMs: A transactive memory perspective on AI literacy [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.26522

Jain, Y., Do, H., Wu, Z., & Wang, A. Y. (2026). A plan-tracing interface for AI-supported algorithm planning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2602.03197

Kao, S., Grant, P., & Woltering, S. (2025). Socratic AI in K–12 science classrooms: Effects on critical thinking, motivation, and self-regulation in a randomized controlled trial [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-8118546/v1

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2506.08872

LearnLM Team, Google, & Fab AI. (2026). Teaching with Gemini: Measuring the impact of Guided Learning on student mathematics progress in Sierra Leone (Technical report). Google. https://storage.googleapis.com/deepmind-media/LearnLM/learnLM_sierraleone_may26.pdf

Lim, W.-C. I., Worden, E., Sales, A., & Heffernan, N. T. (2026). LLM-generated summaries for teachers: A randomized field experiment in a digital learning platform. In Proceedings of the Thirteenth ACM Conference on Learning @ Scale (L@S '26) (pp. 230–241). Association for Computing Machinery. https://doi.org/10.1145/3774398.3811606

Lu, X., Ju, K. P., Dudley, M., Sano, L., & Wang, X. (2026). AI-mediated feedback improves student revisions: A randomized trial with FeedbackWriter in a large undergraduate course [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2602.16820

Mahinpei, R., Dean, V., Fong, R., Liu, L. T., & Horta Ribeiro, M. (2026). AI assistance for discretionary work: Increasing feedback provision in higher education [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2606.03095

Meyer, J., Steinbach, M., Schiller, R., Mertens, U., Schaller, N.-J., Horbach, A., Fleckenstein, J., Köller, O., Kizilcec, R. F., & Jansen, T. (2026). Comparing teacher and AI-generated feedback in the writing classroom: Experimental results from secondary school classrooms. In Proceedings of the Thirteenth ACM Conference on Learning @ Scale (L@S '26) (pp. 264–276). Association for Computing Machinery. https://doi.org/10.1145/3774398.3811602

Morris, C., & Maes, P. (2026). Same feedback, different source: How AI vs. human feedback attribution and credibility shape learner behavior in computing education [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2604.03075

Oreopoulos, P., Liut, M., Sungu, A., & Low, N. (2026). Making AI tutoring productive: Evidence from a mastery-based math practice experiment (NBER Working Paper No. 35621). National Bureau of Economic Research. https://doi.org/10.3386/w35621

Ozturk, A., Aher, G., Mitchell, T., Birgin, M. K., & Kayhan, H. (2026). Who benefits from LLM-generated learning support messages? Causal evidence on treatment effect heterogeneity. In E. G. Blanchard et al. (Eds.), Artificial intelligence in education: AIED 2026 (LNAI 16585, pp. 574–582). Springer. https://doi.org/10.1007/978-3-032-29770-9_62

Rai, A., Chen, L., Ramesh, B., Breazeal, C., Long, Y., Ampel, B., & Aria, A. (2026). Calibrating LLM tutor initiation to learner expertise: Proactive and reactive support for computational thinking [Working paper]. https://bampel.com/uploads/calibrating-llm-tutor-initiation.pdf

Rismanchian, S., Uzun, H., Matayoshi, J., Cosyn, E., & Kurd-Misto, E. (2026). Faster completion, less learning: Generative AI reduced study time on math problems and the knowledge they build [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.21629

Rotter, J., Benazet i Montobbio, P., & Hernández-Leo, D. (2026). Access timing as scaffolding: A reinforcement learning approach to GenAI in education [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.15850

Roy, P., Poet, H., Staunton, R., Aston, K., & Thomas, D. (2024). ChatGPT in lesson preparation: A teacher choices trial (Evaluation Report). National Foundation for Educational Research (NFER). https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/chatgpt_in_lesson_planning_-_evaluation_report.pdf?v=1792501067

Schechter, R., Ackerman, C., Gulemetova, M., & Janakiefski, L. (2025a). Enhancing English language arts achievement through Coursemojo's AI-driven instruction: A quasi-experimental analysis of impact on student outcomes and equity. LXD Research. https://files.eric.ed.gov/fulltext/ED677099.pdf

Schechter, R., Ackerman, C., Gulemetova, M., & Janakiefski, L. (2025b). The impact of Coursemojo AI-powered literacy instruction on sixth-grade English language arts achievement: A quasi-experimental study. LXD Research. https://files.eric.ed.gov/fulltext/ED676253.pdf

Singh, A., Brooks, C., Li, W., Kim, J., & Wang, X. (2026). Hint-writing with deferred AI assistance: Fostering critical engagement in data science education [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2604.19931

Sungu, A., Lira, B., & Duckworth, A. L. (2026). Generative AI can harm teaching [Preprint]. SSRN. https://doi.org/10.2139/ssrn.7007339

Taneja, K., Singh, A., & Goel, A. K. (2026). Impact of multimodal and conversational AI on learning outcomes and experience [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2604.02221

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. (2022). What Works Clearinghouse procedures and standards handbook (Version 5.0). https://ies.ed.gov/ncee/wwc/handbooks

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2025). Tutor CoPilot: A human-AI approach for scaling real-time expertise [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2410.03017

Zhao, C. Q., Cao, J., Lin, J., & Koedinger, K. R. (2026). LLM-based multimodal feedback produces equivalent learning and better student perceptions than educator feedback. In Proceedings of the 16th International Learning Analytics & Knowledge Conference (LAK '26). Association for Computing Machinery. https://doi.org/10.1145/3785022.3785124

Acknowledgments

Project Coordinator: Imogen Lee

Reviewer: Carly Robinson

Images, Design, and Distribution: Julie Brosnan

Funders: Walton Family Foundation, Overdeck Family Foundation, and Google.org

Translating research from a stack of relevant papers to research that is reviewed for rigor and synthesized for meaning requires a team of talented human collaborators. Our Project Coordinator, Imogen Lee, provided the organization and coordination to optimize our final product. Carly Robinson provided key insights into how to shape the content into a coherent narrative. We also thank Julie Brosnan for her creative direction and execution in translating complex data into an accessible and engaging resource.

The AI Hub and the SCALE Initiative are made possible through the Stanford Accelerator for Learning, the Walton Family Foundation, the Overdeck Family Foundation, and with Google.org's support.

A note on AI: This report was written by the authors, built on the outstanding work of our research team, and strengthened by our reviewers. ChatGPT 5.6 Sol, Gemini 3.5, Claude Sonnet 5.5, and Claude Opus 5.5 were used to support text editing. Please see the methodology section for details on AI use in data and research filtering and analysis.

About the SCALE Initiative

The SCALE Initiative is part of the Stanford Accelerator for Learning dedicated to transforming educational opportunity by leveraging knowledge for better education decision-making. SCALE conducts rigorous research, identifies, supports and scales promising solutions and engages decision makers to integrate research, policy, and practice across critical issues in K–12 education.

Learn more at scale.stanford.edu.