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

Lily Fesler JP Martinez Claeys Chris Agnew Susanna Loeb

Stanford | SCALE Initiative
Accelerator for Learning

Executive Summary

This report gives an overview of the characteristics of the current research on AI that is relevant to K–12 settings, and summarizes key takeaways from strong causal studies that examine how AI tools can impact students and teachers.

Research on how AI impacts K-12 students and educators is still extremely limited.

As of October 2025 the AI Hub for Education Research Repository contained over 800 academic papers relevant to AI in K-12 education. Our review found that only a small subset (20 papers) produce strong causal evidence. Causal evidence provides the strongest basis for estimating how a tool impacts students and educators. The current causal research is still very limited: we did not identify any high-quality causal studies in K-12 settings in the U.S. for students and very few for teachers. Much of the research focuses on individuals above 18, is conducted in international settings, is studied in constrained conditions (such as a one-time 20-minute experiment) and focuses on short-term outcomes. However, this small evidence base can still be informative about how the current set of AI tools have impacted students and educators to date.

Research on current AI tools suggests the following key findings:

Students

- **Immediate gains with access:** AI tools significantly improve student performance on math practice, programming projects, and writing tasks while students have active access to the technology.
- **Short-term boost, uncertain transfer:** AI improves performance with access but when assessed independently without AI support, effects are mixed.
- **Easier doesn't mean better:** AI tools can alleviate students' cognitive burden and foster positive experiences in learning, but can be at the expense of deeper thinking.
- **Pedagogical design matters:** Tools designed with pedagogical guardrails (such as AI chatbots for tutoring that provide step-by-step reasoning instead of direct answers) show more promise than general purpose AI tools.

Educators

- **Improved efficiency:** Teachers using AI tools for lesson preparation spent less time on planning without reducing lesson quality.
- **Scaling expertise:** AI tools that provide regular, automated feedback and diagnostics to human tutors can improve instructional quality and student outcomes. AI pedagogical support can be particularly effective for less experienced and lower-rated tutors.

Equity and Student Wellness Challenges

The impact of AI on both educational equity and student emotional and social development remains largely unexamined in the current causal literature. AI tools could reduce achievement gaps if pedagogically sound AI tools provide high-quality, individualized support to students, and if AI tools are effectively used to support less experienced instructors. However, this will depend on whether under-resourced districts have funding for education-specific AI tools. The rapid rise of AI outside of school, including as social companions, raises questions on the impacts for student safety, wellness, and prosocial skill development.

Conclusion

Although the causal research is limited, early evidence still provides useful insights to education leaders about what to consider in incorporating AI tools into their schools. The current evidence suggests that tools designed to foster independent reasoning may be more likely to support durable learning in students. It also suggests that AI tools can help educators complete their tasks more efficiently and improve their pedagogical practices. Given the opportunities and risks, further research is needed to better understand how AI is used, under what conditions, and the impacts for both equity and student wellness. As AI is becoming more prevalent in students' lives both inside and outside of school, understanding how its use shapes learning, human relationships, and developmental outcomes can help educators and policymakers make more informed decisions about how AI is integrated into education and whether its use advances broader education goals.

Table of Contents

I. **Introduction** — **5**

II. **Characteristics of AI Research in Education** — **9**

III. **Key Findings** — **15**

Students — 15

Educators — 24

IV. **Conclusion** — **29**

Appendix — **31**

A. Research Summary Table — 31

B. Research Repository — 33

C. Report Methodology — 35

References — **38**

Acknowledgments — **41**

Recommended citation:

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

Introduction

Teachers, school leaders, and policymakers across K-12 education are navigating a rapidly expanding landscape of artificial intelligence (AI) tools with little rigorous evidence to guide their decisions. AI tools are developing faster than researchers can evaluate them, leaving educators to make high-stakes choices about technology adoption, implementation, and investment with limited evidence of what works, for whom, and under what conditions. Even as definitive research is still emerging, schools are facing growing expectations to prepare students for an AI-integrated world while also responding to declining student performance predating AI and shifting definitions of future readiness.

We launched the AI Hub for Education's Research Repository (the Research Repository) in early 2025 with the goal of helping education leaders identify recent research on AI in K-12 education.

We update the Research Repository monthly as a searchable and filterable resource. The Research Repository contained over 800 papers as of October 2025, and this report draws from that October snapshot. Of those papers, only a small subset rigorously identifies the causal effects of AI tools on students and educators. Using a combination of AI screening and human review, we identified 20 high-quality causal studies in the Research Repository *(see Box 1)*.

This report centers on the current causal evidence on learning outcomes because causal studies are the best way to learn how a tool (like AI) impacts students and educators. Across the limited set of causal evaluations available to date, the first outlines are starting to take shape. Current AI tools often improve student performance and foster positive experiences when students have access to them, but these gains can weaken or disappear when students are assessed without AI

support. For educators, AI tools can save time as well as improve instructional quality. Given the narrow contexts and applications examined in high-quality research to date, these findings should be interpreted as reflecting the limited range of tool uses studied as much as the tools' underlying effects.

This report begins by summarizing the characteristics of the over 800 papers in the Research Repository, which primarily consist of recent preprints (i.e., research papers publicly available but not yet peer reviewed by experts in the field). Understanding what researchers are studying provides insight into where the field is heading, including the topics prioritized, the methods employed, and the outcomes examined. The report then introduces the learning science principles that frame how AI could help or hinder learning opportunities. It then reports overarching takeaways about what is currently known about AI in K-12 education based on the 20 high-quality causal studies.

Overall, this report provides a synthesis of the current causal evidence on AI tools in education, highlighting emerging patterns as well as areas of uncertainty in the literature. In doing so, it aims to provide education leaders with a clearer evidence base for navigating decisions in a rapidly evolving landscape. However, readers should interpret these insights as preliminary and expect them to evolve as more evidence emerges. The field of AI in education is developing rapidly both in products available and insights learned from research. Given this dynamic landscape, many conclusions will require updating as additional rigorous research becomes available. Readers may find it useful to return to the Research Repository regularly as the evidence base develops.

Box 1: Methodology & Considerations

How did we identify papers to analyze in this report?

We considered all papers in the Research Repository as of October 2025. The Research Repository collects papers mostly from arXiv, a free open-access archive where many researchers share their work before submitting it for publication. (*See Appendix B.2)*

What do we consider to be "AI" in this report?

We include papers in K-12 education with any AI or machine learning application that use the terms "AI" or "artificial intelligence." Notably, we do not include studies focused on pre-Large Language Model (LLM) intelligent tutoring systems (often rule-based or probabilistic tutoring platforms) although those tools helped lay the groundwork for today's AI-enabled tools.

What types of studies do the key takeaways come from?

The key takeaways are built upon both preprint and peer-reviewed papers with study designs that support causal inference about the effects of educational interventions, including randomized controlled trials (RCTs) and quasi-experimental designs (QEDs). *(See Appendix C.1)*

How did we review for paper quality?

We identified high-quality causal papers using a two-step quality review process. First, an LLM preliminarily classified papers based on whether they provide strong causal evidence (according to the What Works Clearinghouse [2025]). Human researchers then manually reviewed all prioritized papers for content relevance and the strength of the causal evidence. *(See Appendix C.2)*

How many papers did we identify?

We include the 818 papers in the Research Repository as of October 2025. 20 papers had strong enough causal evidence on educators and students to contribute to the key findings.

Box 2: Report Limitations

What are the limitations of this report?

There are two primary limitations with this report:

1. The Research Repository consists primarily of preprints, which have not undergone peer-review. To mitigate this, at least two human researchers reviewed each of the causal papers discussed in the key takeaways. The report also may not include papers published in journals not covered by our current data collection process.

2. The search query used to identify papers was limited to papers containing the keywords "education" and "AI" or "artificial intelligence." As a result, more recent papers may be more likely to be included in our analysis as the usage of the term "AI" has increased in recent years.

Did we miss a study?

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

Characteristics of AI Research in Education *(to date)*

AI research in education has grown enormously since the release of LLMs a few years ago.[1] In January 2023, only 28 papers had been published that met our Research Repository criteria — AI or machine learning applied to, or relevant for, K-12 education. In less than three years, this number rose to over 800 academic papers (including preprints and journal articles) *(Figure 1)*. In the past year, growth has particularly accelerated, with the number of papers doubling between January and September 2025.

Figure 1: Cumulative number of papers included in the Research Repository

In this section, we provide an overview of the characteristics of the 818 papers in the Research[1] Repository, as well as the 20 high-quality causal impact papers that we discuss more deeply in the

[1] For this report, we define AI research in education to include research that focuses on an AI or machine learning application in K-12 education, and that uses the term "AI" or "artificial intelligence." We do not include studies focusing on pre-LLM intelligent tutoring systems.

next section of this report. We explore who these papers focus on (such as students or educators), the outcomes studied (such as literacy or math), the school level at which these studies were conducted (such as elementary or middle school), and the study designs used (such as RCTs, QEDs, or descriptive methods).[2]

Most AI in education papers focus on students as users.

Over half (59%) of the papers in the Research Repository study students as AI users, and almost three-quarters (70%) of the causal impact papers focus on students *(Figure 2)*. Less than half of papers focus on educators as users (48% of the papers in the Research Repository, and 40% of the causal impact papers). Many papers also examine AI tools used by both teachers and students (23%). Limited research addresses AI use by school leaders and parents and caregivers, representing about 3% of papers in the Research Repository.

Figure 2: Percentage of Research Repository and causal impact papers by the **user of the AI tool**

Note: Studies can focus on multiple users, such as both students and educators. Other Users include school leaders and parents or caregivers.

2 Classification for the characteristics summarized in this section was validated against human coding; for all categories reported here, the automated coding showed reasonable agreement with human coders in our validation checks.

Causal impact papers disproportionately focus on the impacts of AI on math skills.

Although only 17% of papers in the Research Repository study math skills, approximately 35% of the causal impact papers focus on math skills *(Figure 3)*. In contrast, although almost 50% of papers in the Research Repository focus on other academic outcomes (such as science, programming, language, and social studies), only 25% of causal impact papers focus on those outcomes. Additionally, 20% or less of the causal impact papers focus on literacy (20%) and social-emotional outcomes (15%).

Figure 3: Percentage of Research Repository and causal impact papers by the **study outcome**

Note: Literacy includes learning in reading, writing, or language arts skills. Other academic outcomes include learning in science, programming, language, social studies, and other subjects (besides literacy and math). Social-emotional includes improving skills like self-awareness, empathy, and self-regulation. Studies can focus on multiple outcomes, such as both math and other academic outcomes.

Most AI research in education is conducted in postsecondary settings, but the causal evidence base is more focused on high school contexts.[3]

Almost two-thirds (64%) of the research in the Research Repository, compared with roughly 35% of the causal research, is conducted in postsecondary settings. In contrast, causal impact studies are more likely to be conducted in high school settings (45% compared to 38% in the full Research Repository), and less likely to be conducted in middle school and elementary school settings. The emphasis on postsecondary in the broader literature may reflect the fact that AI tools may have been introduced in postsecondary earlier than in K-12 schools and that research is easier to conduct in those settings (it can be easier to recruit, randomize, and collect data from college students who are generally adults).

Figure 4: Percentage of Research Repository and causal impact papers by **level of education**

Note: Studies can focus on multiple education levels, such as both middle school and high school.

[3] Although the Research Repository is focused on findings relevant to K-12 education, it also includes studies conducted in postsecondary settings that could have applications to K-12 education. See Appendix B.2.

Only a small proportion of the research on AI in K-12 is causal.

Most studies in the Research Repository are descriptive (46%) or technical or computational (46%). A much smaller percentage of papers are RCTs (8%) or QEDs (5%). Many papers also use other quantitative methods (30%). This indicates that causal studies remain a relatively small portion of the evidence base. Of the causal studies, 90% are RCTs and around a fifth (20%) are QEDs (some studies have both RCT and QED components).

Figure 5: Percentage of Research Repository and causal impact papers by **study design**

Note: Technical/computational is defined as research focusing on algorithm development, model benchmarking, creation of new datasets, or other computer science-related outcomes. The methodology generally includes computational experiments, including architecture descriptions, training procedures, dataset creation, performance evaluations, or ablation studies. Some examples of papers in this category are Enhancing Retrieval-Augmented Generation with Entity Linking for Educational Platforms and Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability to Mark Short Answer Questions in K-12 Education. Other quantitative methods include mixed-methods studies, observational studies, case studies with quantitative components among others. Some of the causal impact papers include descriptive, technical/computational, or other quantitative methods (in addition to causal methods), which is why those papers are included in the causal impact papers.

Box 3: Learning Science Lens

How can learning science principles help us understand how AI could help or hurt students?

Learning science principles can provide a framework to understand how to interpret key findings based on how students learn best. The below table outlines some key learning science principles and corresponding AI opportunities and risks.

Learning Principle	Description	AI Opportunity and Risk
Cognitive Load Theory (Sweller, 1988)	Manages the limited capacity of working memory by balancing intrinsic, extraneous, and germane (productive) loads.	AI can reduce extraneous load by efficiently retrieving and organizing information, potentially freeing cognitive resources for deeper learning, but it can also reduce germane load — the productive struggle essential for learning.
Vygotsky's Zone of Proximal Development (Vygotsky, 1978)	The optimal learning zone between what a learner can do independently and what they can achieve with appropriate support *(Vygotsky, 1978)*.	The most effective AI tools would provide scaffolds within this zone and gradually release responsibility to the learner to prevent student dependency.
Transfer of learning	The process of applying knowledge gained in one context to new situations, which often requires explicit instructional support connecting the contexts *(Barnett & Ceci, 2002)*.	One key question is whether practice with AI tools develops durable knowledge and skills that students can apply in new contexts, or whether it creates tool-dependent performance.
Metacognition	The ability of students to monitor their understanding, identify gaps, select appropriate strategies, and adjust their approach based on feedback.	Metacognition is difficult to measure and AI could measure metacognition at scale. At the same time, when AI tools perform complete tasks for students, opportunities to develop metacognitive skills may be reduced.
The expertise reversal effect (Kalyuga, 2007)	The phenomenon where instructional techniques effective for novices (such as worked examples) become ineffective or even counterproductive for more advanced learners (who may benefit more from independent problem solving).	Effective AI tools would adapt their support level to learner expertise.
Desirable difficulties (Bjork, 1994; Bjork & Bjork, 2011)	Challenges during practice that produce better long-term retention and transfer, even though they feel less effective and produce lower immediate performance *(Bjork, 1994; Bjork & Bjork, 2011)*.	AI tools would ideally introduce appropriate desirable difficulties, even if users prefer easier practice sessions.

Key Findings

Students

Students engage in a wide range of learning activities inside and outside the classroom, and often require substantial support to learn content and develop key competencies. AI tools are increasingly positioned as scalable solutions that can potentially support these learning processes by offering students on-demand explanations, feedback, and guidance. Most of the tools examined in causal studies are 1:1 chatbots. Beyond chatbots, there are a variety of ways AI could be used to improve students' learning experiences that causal studies have not yet examined, including incorporating the social aspects of learning, developing assessment applications, and supporting diverse learning needs, such as those of English language learners.

Fourteen causal studies related to student use of AI met our quality standard.[4] These papers cover two topics:

- How AI tools affect student performance on practice problems, projects, and assessments
- How different features of AI tools can have differential effects on student performance

These papers also cover two different age groups:

- Seven experiments conducted in universities (three in the U.S. and four in Germany)
- Seven studies conducted with high school students (one in Turkey, two in the UK, two in Germany, one in Belgium and Spain, and one in Brazil), including six experiments and one QED

[4] See Appendix A for a summary table of all studies discussed in this section.

None of these student-facing causal studies were conducted in U.S. K-12 school settings. Understanding how current AI tools impact student learning and performance in U.S. K-12 classrooms — with attention to differences across use cases, subjects, student populations, and school contexts — will require substantially more research. The current evidence offers early insights, but context-specific research in U.S. settings doesn't yet exist.

While international studies provide valuable signals about AI's potential effects, differences across educational contexts may limit how directly findings generalize to U.S. schools. Curricula, instructional practices, student populations, and technology infrastructure vary substantially across countries, and these contextual factors may shape both how AI tools are used and how they influence learning outcomes. As a result, the current evidence should be interpreted as suggestive rather than definitive with respect to U.S. K-12 settings.

Despite these limitations, certain findings are more likely to transfer across contexts. In particular, evidence about how tool design features affect learning processes (such as the difference between general-purpose and tutoring-specific chatbots, or the impact of reduced cognitive load on reasoning quality) likely reflects fundamental cognitive principles rather than context-specific factors.

1 Some student-facing AI tools improve performance while students have access, but effects when students no longer have tool access are mixed.

Studies to date have generally found that AI tools can improve student performance while they have access to those tools. AI tools (including automated feedback tools and general-purpose and tutoring AI chatbots) have improved student performance on practicing math proofs for college students in the U.S. (Chen et al., 2025; Zhao et al., 2025), math practice problems for high schoolers in Turkey (Bastani et al., 2025), an economics exam for students in Germany (Fischer et al., 2025), writing an argumentative essay for students studying English as a foreign language Meyer et al., 2024), and a physics question for high-school-aged students in Spain and Belgium (Blasco & Charisi, 2025). However, one experiment with college students in Germany found that students using general-purpose AI chatbots to conduct research demonstrated lower-quality reasoning and argumentation compared to those using a traditional search engine (Stadler et al., 2024).

In contrast, studies have found mixed impacts of students' practicing with AI tools to study for assessments they take without AI access. One study found that giving high school students in Brazil AI-generated feedback on their essays improved their scores on a high-stakes argumentative writing exam (Ferman et al., 2021). However, one experiment in the U.S. found that giving college students access to automated feedback and a tutoring-specific AI chatbot to study for a math exam did not improve their exam scores (Chen et al., 2025). An experiment in Turkey found that high school students who practiced for an exam using a AI chatbots either performed worse on the final exam (if they used the general-purpose AI chatbot) or the same (if they used a tutoring-specific AI chatbot) compared to their peers (Bastani et al., 2025). Another set of studies found mixed effects of using general-purpose

Learning Science Lens

When AI tools perform information retrieval, organization, and initial processing for students, opportunities to develop **metacognition** may be reduced. Students may have fewer occasions to evaluate source credibility, identify gaps in their understanding, or select appropriate problem-solving strategies if the AI performs these functions for them.

AI chatbots to learn programming, with a couple of lab experiments showing no effects of AI chatbots and a field study indicating worse learning outcomes from using a general-purpose AI chatbot to study (Lehmann et al., 2025).

Voices of the Research

> "While AI can reduce perceived difficulty and increase fluency, this may come at the cost of reduced independent reasoning and weaker knowledge acquisition."
>
> *– Dr. Matthias Stadler, University of Munich*

These findings provide evidence of an important distinction between tool-supported performance and durable learning. AI tools may help students complete tasks more successfully in the moment, but those gains do not always persist when students are later asked to perform independently. One possible explanation is that students may be learning how to work with the tool rather than developing the underlying knowledge and reasoning skills needed for unaided performance. For example, Bastani et al. (2025) found that students using a general-purpose AI chatbot performed substantially worse than students with no access to AI chatbots on a closed-book exam, even though they performed better during AI-supported practice. Stadler et al. (2024) similarly found that students using a general-purpose AI chatbot experienced lower mental effort and produced weaker reasoning and argumentation than peers using traditional search.

A second possibility is that students may be learning content, but in ways that do not transfer when the context or available supports change. Transfer is difficult to achieve even under traditional instruction, and AI tools under studied contexts may further complicate whether students internalize skills in a form that supports flexible application. Consistent with this concern, Chen et al. (2025) found that automated feedback paired with an AI chatbot

Learning Science Lens

AI tools that improve performance when students have access to the tool but not when the tool is removed suggest that students may not be **transferring learning** (developing durable knowledge and skills that students can apply in new contexts).

for tutoring improved homework performance but did not improve exam scores when AI was no longer available.

These findings suggest that how AI tools are designed and used may be an important factor in effectiveness. Tools that provide structured guidance, hints, or step-by-step reasoning may be more likely to support durable learning than tools that provide complete solutions. More research could clarify when AI serves as scaffolding that strengthens both student learning and social engagement, versus when it risks reducing the cognitive work required for long-term mastery.

2 AI tools can alleviate students' cognitive burden and foster positive experiences in learning, although they may not encourage deeper thinking.

Students experience significant cognitive relief during challenging tasks, such as solving mathematical proofs, programming, or conducting scientific inquiries. As a result, they express greater enjoyment in their learning experiences (Becker et al., 2025; Chen et al., 2025; Stadler et al., 2024; Zhao et al., 2025).

Although the freed-up cognitive resources may enhance some aspects of learning, they do not necessarily result in deeper learning, especially in higher-order cognitive tasks like reasoning and argumentation. Research indicates that students who utilized

Learning Science Lens

Some AI tools may reduce productive cognitive effort (or germane load) in addition to reducing unnecessary cognitive effort (extraneous load) (which are both types of **cognitive loads**). Students may prefer AI tools that make learning easier, but engaging in cognitively challenging tasks (known as "**desirable difficulties**") often produces better long-term retention and transfer.

Voices of the Research

> "For future research we are interested in exploring how we use these tools to effectively create productive struggle to enable learning. Students keep bypassing heuristic guardrails, can we build a "science of guardrails" to understand what works in education?"
>
> *– Dr. Hamsa Bastani, University of Pennsylvania*

general-purpose AI chatbots generated lower-quality reasoning and argumentation in their analyses compared to those who used traditional search engines (Stadler et al., 2024). Other studies suggest usage of general-purpose AI chatbots to help with a writing task reduced brain activity and led to weaker recall (Kosmyna et al., 2025). Research on reading comprehension found high-school-aged students perceived AI use to be more enjoyable and helpful than traditional learning methods, but their retention improved when AI use was complemented by traditional learning strategies like notetaking (Kreijkes et al., 2026).

3 Tutoring chatbots may be more effective than general-purpose AI chatbots, and the pedagogical design of the tutoring chatbot matters.

An experiment in Turkey found that students who had access to a general-purpose AI chatbot to study for an exam performed worse than their peers who worked through practice problems in a course textbook, but that students who instead had access to a tutoring-specific AI chatbot performed the same as their peers working in textbooks (Bastani et al., 2025). In that study, the tutoring-specific AI tool gave hints to the student without directly giving the answer.

A couple of studies also examined Socratic AI chatbots (that ask users probing questions instead of providing a direct answer) and found mixed results. A small, lab-based randomized study of high-school-aged students found that a Socratic-style AI chatbot that led students through layers of guided questioning was rated as less helpful by students compared to a chatbot that gave students the answers directly (Blasco & Charisi, 2025). Another study of pre-service teacher education students in Germany rated a Socratic AI chatbot as more supportive for critical, independent, and reflective thinking than a general-purpose AI chatbot (Degen et al. 2025). Taken together, this could mean that students often prefer AI chatbots that provide direct answers to Socratic chatbots and prefer Socratic chatbots to general-purpose chatbots. Alternatively, it could mean that students' preferences for Socratic chatbots depend on the context and the specifics of the tool.

Learning Science Lens

General-purpose AI tools that provide complete solutions may operate outside students' **Zone of Proximal Development (ZPD)**, either doing the work for them (requiring no cognitive stretch) or providing information they cannot yet make sense of. Tutoring-specific tools that provide hints and step-by-step reasoning may better target the ZPD by offering graduated support.

Equity Considerations for Student-facing AI Tools

The current evidence base provides limited insight into how AI tools affect educational equity. AI tools have the potential to provide individualized academic support at scale, which could benefit students who lack access to private tutoring or other supplemental resources. Students with learning needs, English language learners, and students with disabilities could all potentially benefit from AI support in the right context.

Equity in access and use depends on multiple factors beyond whether a tool exists. Students' ability to benefit from AI tools may vary with technology infrastructure, digital literacy, and whether students can access tools at home as well as at school. Language accessibility is also a key concern, as many tools are optimized for English and may provide lower-quality or biased support for English Learners. Similarly, AI tools could offer new accommodations for students with disabilities, but the current research does not examine impacts for students with Individualized Education Programs (IEPs) or 504 plans.

These gaps leave several equity questions unanswered. Do AI tools disproportionately benefit students who already have stronger academic preparation and support outside of school, or can they help level the playing field for students with fewer resources? How do costs and licensing models shape which schools can access higher-quality tools? Few studies examine equity effects in the current causal literature.

Student Wellness Considerations for Student-facing AI Tools

Unlike previous education technologies, student-facing AI blurs the line between general-purpose technology and purpose-built education tools in ways that have implications beyond the classroom. Like social media, AI use is not isolated to supervised school settings. Survey data suggests a rapid increase of AI tool use that seeks to replicate human relationships and cultivate emotional bonds and personal rapport with users, including children and teens (Robb & Mann, 2025). This raises important questions about the impacts of AI use outside of school on students' emotional, social, and cognitive development.

Our review finds limited causal evidence on AI's effects on both cognitive development and student emotional or social wellness. The gap highlights important unanswered questions for K-12 practitioners and policymakers: What conditions support prosocial development when students interact with AI? What are the effects of AI social companions on children and adolescents? And what practices are most effective for promoting safe AI use outside of school, on both personal and school-owned devices?

The SAFE AI Companions Task Force, convened by the EDSAFE AI Alliance (2026), developed a research agenda on AI companions in education organized around five key topics: mandated reporting, student data privacy, prosocial design and use, learning sciences and effective pedagogy, and benchmarking. As the experience of social media has shown, causal research on technology's effects on student wellness is difficult to conduct and the stakes are high. Informed decision making and sound policy depend on building this evidence base.

Connecting Student- and Educator-facing Findings

Across the studies reviewed, student-facing AI tools (typically in the form of a 1:1 chatbot) often improve performance while students have access to support, but effects on independent learning and transfer are less consistent. This pattern highlights the distinction between tools that support short-run task completion and those that strengthen durable learning.

Educator-facing tools, in contrast, are more often designed to augment instruction by helping teachers target attention, provide feedback, or allocate instructional time. Across both domains, the evidence points to the importance of tool design and human mediation in shaping outcomes.

Educators

Educators are central to creating effective learning environments, but face significant demands on their time and expertise as they work to meet diverse student needs. Educators are increasingly exploring AI tools to automate routine tasks, provide data-driven insights about student progress, and offer suggestions for instructional approaches. Beyond efficiency gains, AI may enable educators to personalize their support more effectively, devote more attention to students who need it most, and receive real-time coaching to refine their teaching practice. As with student-facing applications, the impact of AI on educators depends on how these tools are designed and integrated into existing workflows and educational contexts.

Eight causal studies related to educator use of AI met our quality standard.[5] These papers cover several topics:

- How AI affects teacher time use
- How AI affects the quality of teacher outputs (such as lesson planning and student feedback)
- How AI can be used to provide real-time feedback to teachers and tutors in remote and face-to-face settings
- The types of teachers who benefit the most from AI support

These papers also cover two different settings:

- Four experiments in K-12 settings in the U.S.
- Four experiments in K-12 settings in international contexts (including in Brazil, England, and South Korea)

[5] See Appendix A for a summary table of all studies discussed in this section.

1 AI can reduce teacher time spent on routine tasks, or shift teacher effort, with no evidence of quality losses.

Teachers given access to ChatGPT and a guide on how to use it spent about 30% less time on lesson and resource preparation (about 25 minutes per week), with no detectable differences in lesson quality based on blind expert ratings (Roy et al., 2024). In contrast, an AI-only automated writing evaluation system did not reduce teachers' total work-from-home hours, even as student writing outcomes improved, suggesting that teachers may have reinvested time saved on routine feedback into higher-skill instructional support rather than reducing total hours worked (Ferman et al., 2021). This same study found that teachers who had access to AI automated writing feedback tools discussed more essays individually with students.

Research also provides evidence that teachers can begin saving time with AI tools quickly and become more efficient in how they use them over time. Teachers with access to ChatGPT spent 27% less time creating lessons than their peers in the first five weeks of use, and 31% less time in weeks six through 10 (Roy et al., 2024). The study also found that teachers' use of AI tools decreased over time (from 39% to 29% of total lessons generated) yet savings persisted, suggesting teachers quickly learn where AI adds value and deploy it more selectively.

Voices of the Research

> "In the future, I would like to see more [research] about how teachers can benefit from AI while maintaining quality of lessons and teaching. Accuracy of AI generated content when teaching is clearly vital, so greater understanding of how less experienced teachers can benefit from this without having to spend extra time checking would be useful."

– Dr. Helen Poet, National Foundation for Educational Research

2 AI can deliver automated feedback on teaching and student progress, improving instructional quality and student outcomes.

Weekly AI reports analyzing classroom discourse have been shown to improve pedagogical practices, such as increasing teachers' use of "focusing questions" by 20% in brick-and-mortar classrooms (Demszky et al. 2025). Similar automated feedback tools in large-scale online courses have improved instructors' "uptake" of student ideas by 10%, improving student course satisfaction (Demszky et al. 2023).

Furthermore, providing weekly diagnostic reports to tutors — detailing student performance patterns, time allocation, and areas of difficulty — have yielded significant improvements in student learning outcomes on assessments (Kim et al., 2021).

3 AI can offer real-time instructional suggestions that improve educator practice and student outcomes, particularly in messaging-based settings.

In remote tutoring settings, AI can provide real-time, context-specific suggestions that expand tutors' pedagogical strategies, enabling them to deploy a broader range of teaching strategies during instruction than they would typically employ on their own (Wang et al., 2025; LearnLM Team, Google & Eedi, 2025). For instance, the Tutor CoPilot study demonstrated that providing real-time suggestions based on expert reasoning encouraged tutors to use strategies like asking guiding questions more frequently. Tutors who used LearnLM also reported in surveys that they appreciated LearnLM's strength at drafting Socratic questions that tutors could send to students. Both programs also improved student outcomes: Tutor CoPilot improved student topic mastery by 4 percentage points, and LearnLM increased student performance on subsequent topics by 5.5 percentage points.

Emerging studies show some promise in using novel technologies to facilitate real-time feedback in in-person or voice-based interactions, but more research could illuminate the most productive approaches. In one study, teachers were given smart glasses that provided a virtual layer of real-time analytics identifying students who were struggling with tasks (Holstein et al., 2018). This technology enabled teachers to allocate more instructional time to students with lower prior skills and promote greater learning gains among struggling students. Additional research could help clarify how current AI tools shape student learning and performance in U.S. K-12 classrooms, including variation across subjects, student populations, and school contexts.

4 AI pedagogical supports appear to be most beneficial to lower-rated and less experienced educators.

AI support, through real-time, expert-like suggestions for instructors provided by systems like Tutor CoPilot, appears to be particularly beneficial for tutors with lower ratings and less experience (Wang et al., 2025). In a large-scale RCT with 900 tutors, students whose tutors used the system were 4 percentage points more likely to master lesson topics, a benefit that increased to 7 percentage points for the students of tutors with less experience and 9 percentage points for the students of lower-rated tutors. These findings are notable given that traditional professional development opportunities — particularly time-intensive supports such as classroom observations or coaching from veteran instructors — are often limited in availability. AI-generated support can quickly provide inexperienced educators with actionable feedback that develops their pedagogical skills and helps them address student learning gaps.

Learning Science Lens

Expertise reversal effect: The effectiveness of AI supports may depend on learner expertise levels. Beginners may benefit from more explicit AI guidance, while advanced learners may benefit more from minimal AI intervention. This raises questions about whether AI adapts its support level to learner expertise, providing more scaffolding for novices and less for experts, and how effects vary by student prior knowledge.

Equity Considerations for Educator-facing AI Tools

AI tools for educators raise distinct equity considerations related to teacher quality distribution and access to professional learning. One potentially important finding in the current evidence is that AI pedagogical supports may be especially beneficial for tutors with lower ratings and less experience (Wang et al., 2025). If AI tools can help less experienced educators improve their practice, they could help address persistent inequities in access to effective instruction, since under-resourced schools disproportionately employ novice teachers.

At the same time, equitable access to educator-facing AI tools is not guaranteed. Under-resourced districts may lack funding for licensing education-specific tools, leading teachers to rely on free general-purpose systems that may be less effective or raise privacy concerns. Teachers also need time, training, and infrastructure to integrate AI tools into their work, and these supports are often unevenly distributed across schools.

The current research does not yet clarify whether AI tools help educators build lasting instructional skills or create new forms of dependence on automated guidance. Additional evidence could clarify how AI affects teacher practice and student outcomes across different school contexts, and whether these tools reduce or reinforce existing disparities in instructional quality.

Voices of the Research

> “A lot of tools are trying to automate teacher tasks, which can be helpful in terms of efficiency, but could lead to skill degradation and also does not guarantee good instruction quality [since] the models are far from being able to reproduce excellent teacher practice.
>
> We instead are prioritizing educative **tools that seek to build teachers' capacity by supporting their own learning.** AI can be very helpful in this space given that traditional, human-based professional learning can be very costly (e.g. coaching), and/or ineffective (one-size fits all modules), and AI can provide customized, consistent feedback to teachers at a low cost. This can also complement human coaching.”

– Dr. Dora Demszky, Stanford University

Conclusion

The evidence reviewed in this report provides early insight into how AI tools may shape teaching and learning, while also making clear how limited the current knowledge base remains. The research is still narrow in scope: most causal studies focus on math and computer science, with less evidence in literacy, social studies, language learning, and other core subject areas. Many interventions examined were short in duration and relied on within-software learning measures rather than external assessments, leaving open questions about whether observed effects transfer to broader academic outcomes or persist over time. In addition, much of the existing evidence comes from studies conducted outside U.S. K-12 settings, which may limit direct applicability to American schools and classrooms.

The studies available so far also provide only partial insight into what aspects of AI tools matter most for student learning. Much of the literature compares AI access to no AI access, leaving open questions about how different design choices, levels of guidance, or interaction structures shape learning processes and outcomes. Additionally, most student use was in the form of 1:1 chatbots with little to no exploration into AI use that promotes collaboration and human interaction in learning. The evidence base remains concentrated in a narrow set of outcomes, particularly short-term performance measures, with less known about effects on engagement, metacognition, self-regulated learning, or other longer-term competencies. Existing studies further suggest that impacts may vary substantially across learners, raising questions about which students benefit most or least from AI tools and under what conditions. More research is also needed on how AI can be effectively integrated into teacher practice. Lastly, there exists no high-quality causal research on the timely topic of developing student and teacher AI literacy.

The pace and type of AI development will likely shape how evidence accumulates over time. Studying AI's impacts on social learning, motivation, and human augmentation remains an important priority. Understanding AI's effects on longer-term outcomes, such as skill development,

graduation, postsecondary persistence, or labor market success, will likely require multi-year longitudinal research. Similarly, assessing whether AI tools alter students' cognitive development or capacity for independent problem-solving will require longer time frames. Challenging this, the rapid evolution of AI systems means that findings may be closely tied to particular tool versions and implementation contexts. This creates inherent tension between the value of longer-term studies assessing sustained impacts and shorter-term studies that provide timely evidence about tools currently in use. Over time, a clearer understanding of AI's educational role will depend not only on whether tools appear effective on average, but also on when, how, and for whom they matter most. Answering these questions will be central to move from early, mixed findings toward clearer conclusions about impact.

Appendix A
Research Summary Table

Legend

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

The 20 papers with strong enough causal evidence on educators and/or students to contribute to the key findings.

Paper	User	Intervention Description	Sample	Key Finding
Kreijkes et al. (2026)	Students	Students studied two text passages using a GPT-3.5 chatbot, traditional note-taking, or a combination of both, following instructions for active, comprehension-focused reading.	344 Year 10 students (ages 14-15) in England	Note-taking (alone or with an AI chatbot) significantly outperformed AI-only use for comprehension and retention. Students preferred the AI chatbot and perceived it as more helpful, despite it leading to the lowest learning performance.
Fischer et al. (2025)	Students	Students used a GPT-4 based AI chatbot (with Retrieved Augmented Generation) to study economics material, comparing restricted access (10min delay) vs. unrestricted access vs. no access (textbook only) in a 25min session.	334 university students in Germany	Students with access to the AI chatbot outperformed students with no access to AI, and students with unrestricted access outperformed students with restricted access.
Bastani et al. (2025)	Students	Students engaged in GPT-4 math practice with pedagogical guardrails (GPT Tutor) vs. a standard interface (GPT Base). Students participated in four 90-minute sessions over one semester.	~1,000 high school students in Turkey	Unfettered AI access improved practice grades but led to a 17% performance drop on unassisted exams; guardrails (GPT Tutor) mitigated this "crutch" effect.
LearnLM Team, Google & Eedi (2025)	Students, Educators	Students chatted with LearnLM (an AI tool fine-tuned for pedagogy and integrated into math tutoring chats and supervised by expert tutors) vs. chatted with human tutors. The intervention was over seven consecutive weeks.	165 Year 9 /10 (ages 13-15) students across 5 schools in the UK	Students with AI-guided tutors performed at least as well as those with human tutors and showed superior knowledge transfer to novel topics.
Kosmyna et al. (2025)	Students	Participants wrote essays using AI chatbot assistance (ChatGPT), search engines, or no tools. A subsequent condition swap tested the effects of removing AI support.	55 undergraduates, postgraduates, and university staff.	Writing essays using an AI chatbot reduced recall, with 83% of participants failing to be able to provide a quote from their essay, compared to 11% of participants who used a search engine or no tools.
Blasco & Charisi (2025)	Students	Students interacted with an AI chatbot that gave step by step reasoning vs. an AI chatbot that gave the solution alone, and a Socratic AI chatbot vs. a direct AI chatbot. Users completed one 60-minute session with the AI chatbot.	122 high school students in Belgium and Spain	The AI chatbot that provided step-by-step reasoning significantly improved performance compared to receiving the solution. The Socratic chatbot increased engagement but was perceived as less helpful than the direct chatbot.
Becker et al. (2025)	Students	Students used a custom-configured AI chatbot vs. tiered hints vs. a digital textbook page to solve a physics problem.	273 9th grade students in Germany	Both the AI chatbot and tiered hints significantly reduced cognitive load and improved affective outcomes (such as enjoyment and hope) compared to the digital textbook condition.
Chen et al. (2025)	Students	Students had access to LLM-Tutor (which provided AI feedback on LaTeX proof submissions and a math-themed AI chatbot) vs. no AI feedback. It was a semester-long study.	155 university students in a discrete math course in the U.S.	Early AI access significantly improved homework performance, but had no significant impact on unassisted exam scores.
Lehman et al. (2025)	Students	Students had unrestricted AI chatbot access (ChatGPT 3.5) while learning Python programming vs. no AI access.	Postsecondary students in Germany and the Netherlands.	AI chatbot access had no overall effect on learning; AI use increased the volume of topics covered but harmed student understanding. It also widened achievement gaps for students with low prior knowledge.

Degen et al. (2025)	Students	Students interacted with a Socratic AI Tutor (scaffolded via structured dialogue) vs. an uninstructed chatbot to refine a research question during a 5-minute session.	65 pre-service teacher students in Germany	The Socratic AI chatbot significantly increased perceived support for critical, independent, and reflective thinking compared to the uninstructed AI chatbot.
Meyer et al. (2024)	Students	Students received AI-generated feedback (GPT-3.5-turbo) on an argumentative essay vs. no AI access. Conducted during a 90-minute classroom lesson.	459 upper secondary students (Grade 10) in Germany	AI-generated feedback significantly increased students' essay revisions, task motivation, and positive emotions compared to not receiving feedback.
Zhao et al. (2025)	Students	Students received AI autograder feedback for mathematical proof-writing vs. no AI access. Students completed one 60-minute assessment over the course of one week.	169 university students in the U.S.	Students using AI autograder feedback achieved significantly higher scores (approx. 11 points more) than those doing self-evaluation.
Stadler et al. (2024)	Students	Using ChatGPT-3.5 versus Google search to research scientific issues in a one-time 20-minute session and provide recommendations.	91 university students in Germany	AI chatbots significantly reduced students' cognitive load while they conducted research and reduced the quality of students' reasoning and argumentation compared to using a traditional search engine.
Ferman et al. (2021)	Students, Educators	Students received writing feedback from only AI vs. AI and independent human graders vs. not receiving any writing feedback. The program was implemented over one academic year.	Approximately 19,000 public school students in Brazil across 178 schools	Receiving AI feedback significantly improved student essay scores compared to the control group; the addition of human graders did not further improve performance despite students perceiving the feedback to be higher quality.
Demszky et al. (2025)	Educators	TeachFX delivered automated feedback to teachers via a mobile app and weekly emails that targeted "focusing questions" (to probe student thinking) vs. no feedback. The intervention took place in-person over five months.	369 math and science teachers in the U.S.	Automated feedback increased teachers' use of focusing questions by 20%. There was no discernible effect on student talk time or reasoning.
Wang et al. (2025)	Educators	Tutor CoPilot gave real-time, expert-like AI suggestions to human tutors during math sessions vs. no suggestions. Students received tutoring over the course of two months.	900 tutors and 1,800 K-12 students from Title I schools in the U.S.	Access to Tutor CoPilot significantly improved student mastery of lesson topics (4-14 p.p.), and the benefits were largest for students of lower-rated and less-experienced tutors.
Roy et al. (2024)	Educators	ChatGPT for lesson preparation (using a specific guide) vs. non-AI preparation. Teachers in the treatment group had access to ChatGPT for ten weeks and were provided with a guide on teaching with ChatGPT.	259 Year 7/8 science teachers across 68 schools in England	ChatGPT saved teachers an average of 25 minutes per week (31% reduction) on lesson preparation while maintaining the same quality.
Demszky et al. (2023)	Educators	M-Powering Teachers (automated NLP-based feedback to teachers on their "uptake" of student ideas) vs. no feedback. Access to the automated feedback was provided over 10 sessions.	414 instructors from a U.S.-based online platform for high school students	Automated feedback improved instructors' uptake of student ideas by 10%, reduced instructor's talk time by 5% and improved students' experience with the program.
Kim et al. (2021)	Educators	Providing AI-generated weekly student progress/achievement reports to home tutors vs. no reports. Tutors in the treatment group received AI-generated reports for 12 weeks.	234 tutors and 2,220 students in South Korea	AI reports significantly improved student academic performance, though "technology overload" hindered the benefit for certain tutors.
Holstein et al. (2018)	Educators	Access to mixed-reality smart glasses providing real-time AI-generated student analytics to teachers vs. no access to analytics. Students worked with a tutor for a total of 60 minutes, spread across two classes over one week.	286 middle school students and 8 teachers in the U.S.	Smart glasses allowed teachers to redirect attention to struggling students, narrowing the learning gap across prior performance levels.

Appendix B
Background on the Research Repository

This section gives a description of the AI Hub Research Repository and a brief description of the process for paper extraction, eligibility screening, and classification for the Research Repository.

B.1 Overview of the Research Repository

The AI Hub for Education Research Repository is a comprehensive collection of academic research on AI in K-12 education.

The Research Repository aims to include all research on AI in K-12 education in the U.S. As research diverges from AI for K-12 in the U.S. — such as education systems beyond K-12, or studies conducted outside the U.S. — inclusion in the Research Repository is based on relevance to the AI Hub for Education target audiences:

- Superintendents, state, and federal K-12 leaders
- Education support organizations (unions, parent groups, etc.)
- Leadership and product teams at technology companies
- Academic researchers
- Global education leaders

The Research Repository includes academic papers and it does not include news articles on AI in education.

B.2 How papers are collected and organized within the Research Repository

The inclusion of papers into the Research Repository consists of three steps (i) extracting papers related to AI in Education; (ii) filtering papers based on their relevance to AI in K-12 Education in the U.S.; and (iii) tagging papers across different categories to organize papers in a way that is useful for users.

Paper extraction:

Papers included in the Research Repository come from three sources: (i) preprint papers retrieved from arXiv using the arXiv API, (ii) papers suggested by SCALE staff, and (iii) papers submitted by users. We focused on arXiv because researchers often post working papers and preprints there to share findings quickly, often before publication in a peer-reviewed journal, and this speed is especially important given the rapid pace of change in AI. As of October 2025, 87% of the papers in the Research Repository were extracted from arXiv.

For the extraction, a keyword-based query is used in the arXiv API to retrieve papers that contain terms related to AI and education. At any update, papers are retrieved by date, including any paper that has been newly published or updated since the last update through the day of extraction.

Eligibility:
Given that the previous step captures a broad number of papers, the next step is to determine whether an extracted paper is relevant to the Research Repository. Using a prompt-based classifier powered by a large language model ("**LLM**"), each paper is scored on a 1-10 scale based on its relevance to the usage of AI in U.S. K-12 education. Each score is then reviewed by a human researcher (who confirms or modifies the score), and papers with a final score of 8 or higher are included in the Research Repository.

As of March 2026, the papers in the Research Repository are not evaluated for methodological quality.

Classification:
After a paper is determined to be eligible, it is classified using a prompt into one or more values of the following five categories:

1. **Intended application of AI:** Teaching - Instructional Material, Teaching - Assessment and Feedback, Teaching - Professional Support, Learning - Student Support, Communicating / Social Skills, Organizing, and Analyzing.
2. **User role:** Student, educators, school leaders, parents/caregivers, and others.
3. **Student age group affected**: 0-3 years, elementary school, middle school, high school, postsecondary and adult.
4. **Goal of the AI use:** Efficiency, Literacy, Math, Outcomes - Other Academic, Outcomes - Differentiation, Outcomes - Social Emotional, Outcomes - Durable Skills and Reimagining School.
5. **Study design:** Randomized Controlled Trial, Quasi-Experimental, Quantitative - Others, Qualitative - Product Development, Qualitative - Design and Implementations, Technical Computational and Systematic Review.

The value for these categories serve as a filter for users on retrieving papers with specific users, applications, etc.

Appendix C
Report Methodology

As explained in the main section of the Report, this analysis draws on studies listed in the Research Repository as of October 2025. The report focuses on evaluations using **Randomized Controlled Trials (RCTs)** and **Quasi-Experimental Designs (QEDs)**, that were considered to be producing causal evidence.

This appendix explains: (a) the rationale for only including RCT and QEDs, and (b) the process used to extract and synthesize insights from the academic literature.

C.1 Rationale for focusing on RCTs and QEDs

We consider only RCTs and QED studies in our Key Findings because these research designs are best suited to support causal inference about the effects of educational interventions. RCTs establish causal relationships through random assignment, while well-designed QEDs use rigorous comparison groups and statistical controls to reduce selection bias when randomization is not feasible. RCTs are the gold standard for making causal claims, but both designs are capable of producing credible estimates of intervention effects when they meet established methodological standards.

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 design studies (QED) are studies that evaluate an intervention **without random assignment**, but still use a carefully selected comparison group to estimate its effects. Researchers use design strategies and statistical methods to their best of their ability to estimate a causal effect, allowing for reasonable conclusions about whether the intervention contributed to observed differences in outcomes, even when randomization is not practical or ethical.

We do not include descriptive studies, systematic reviews, computational/technical papers, qualitative research, or other non-causal quantitative designs in this causal-evidence section because they do not produce causal evidence (the focus of this report). However, studies with non-causal methods can be highly valuable for understanding AI in K-12 contexts, and can:

- Clarify mechanisms and implementation (how and why an AI tool works, for whom, and under what conditions), which causal estimates alone often cannot explain.

- Surface contextual factors (school constraints, teacher uptake, infrastructure, language access, privacy policies) that determine whether impacts are likely to replicate.
- Identify risks and unintended consequences (bias, overreliance, data security, equity impacts) that may not appear in short-term outcome measures.
- Inform design and measurement by defining constructs, proposing theory of change, and improving outcome instruments used in later RCTs/QEDs.
- Support external validity by documenting settings and populations underrepresented in causal studies.

For this reason, we include these studies in the Research Repository, even though these studies are outside the scope of this report.

C.2 Translating the papers in the Research Repository into findings

After identifying the papers of the Research Repository that would be eligible based on the users of AI and their methodology, we conducted a structured review of the studies in each group to extract insights about how AI is being used by students and educators. Because the report aims to summarize evidence about impact, we prioritized studies with sufficient methodological rigor to support credible causal conclusions. This approach helps ensure that the findings highlighted are not only relevant, but also reliable and interpretable in terms of what AI-based educational interventions appear to do in practice.

To determine eligibility and assess rigor, we followed guidance from the What Works Clearinghouse Procedures and Standards Handbook, Version 5.0 (2025) ("WWC Handbook"). The WWC provides guidance for evaluating whether education studies provide causal evidence, assigning ratings such as Meets WWC Standards Without Reservations, Meets WWC Standards With Reservations, or Does Not Meet WWC Standards. At least one researcher reviewed each paper to assess eligibility. Papers were included for consideration in the Key Findings if considered appropriate in terms of methodological rigor using guidance from the WWC Handbook.

To facilitate review by researchers, we used an LLM to classify papers along two dimensions in two different classification process:

- *Basic alignment criteria:* Drawing on Section 2 of the WWC Handbook, this step checks whether the study's population, intervention, and outcomes are clearly defined and aligned with an evaluation intended to estimate the effects of an intervention.
    - We assessed papers for (1) completeness (i.e., sufficient information to review), (2) whether the population aligns with K-12 education, (3) whether the intervention is relevant to K-12 education, and (4) whether at least one educationally relevant outcome is reported (academic, behavioral, social-emotional, teacher, or school leader outcomes). A paper passed the basic alignment criteria

only if all items were coded True; 93 of 103 papers passed this criterion.

- *Methodology screening:* 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 be conservative in excluding papers, a study did not pass this step if it received a 1 on any dimension; 61 of 103 papers passed this stage.[1]

After classification, researchers responsible for paper review received a list of papers along with each study's pass/fail status for the two criteria. To generate insights from the literature, reviewers were encouraged to prioritize studies that passed both criteria, while remaining free to review additional papers as needed.

Researchers then reviewed the studies to develop the insights presented in the Key Findings Section. Each paper passing both criteria was reviewed by a researcher to confirm relevance to the report and to validate methodological rigor using the WWC Handbook eligibility standards as guidance.

At least one human researcher drafted each key finding which was then verified by a second additional researcher and reviewed by two additional researchers.

[1] Compositional Change and Outcome Reliability were excluded from the evaluation as the results of the classification were not consistent with the quality of papers, with a systematic error towards grading negatively.

References

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

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

Becker, E., Wünsche, J., Veith, J. M., Schrader, J., & Bitzenbauer, P. (2025). From cognitive relief to affective engagement: An empirical comparison of AI chatbots and instructional scaffolding in physics education [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.06254

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185–205). MIT Press. https://doi.org/10.7551/mitpress/4561.001.0001

Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), Psychology and the real world: Essays illustrating fundamental contributions to society (pp. 56–64). Worth Publishers.

Blasco, A., & Charisi, V. (2025). AI chatbots in K-12 education: An experimental study of Socratic vs. non-Socratic AI agents and the role of step-by-step reasoning [Preprint]. SSRN. https://doi.org/10.2139/ssrn.5040921

Chen, E., Judicke, S., Beigh, K., Tang, X., Xiao, Z., Li, C., Li, S., Luttmer, R., Singh, S., Yampolsky, M., Parikh, N., Zhao, Y., Chen, M., Huang, S., Mohanty, A., Johnson, G., Mackey, J., Lin, J., & Koedinger, K. (2025). Generative AI alone may not be enough: Evaluating AI support for learning mathematical proof [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2509.16778

Degen, P.-B., & Asanov, I. (2025). Beyond automation: Socratic AI, epistemic agency, and the implications of the emergence of orchestrated multi-agent learning architectures [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.05116

Demszky, D., & Liu, J. (2023). M-Powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes. In Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23). Association for Computing Machinery. https://doi.org/10.1145/3573051.3593379

Demszky, D., Liu, J., Hill, H. C., Sanghi, S., & Chung, A. (2025). Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers & Education, 227*, Article 105183. https://doi.org/10.1016/j.compedu.2024.105183

EDSAFE AI Alliance. (2026). S.A.F.E. by Design: Policy, research, and practice recommendations for AI companions in education (Task Force report). EDSAFE AI Alliance. https://www.edsafeai.org/safieaichatbots

Ferman, B., Lima, L., & Riva, F. (2021). Artificial intelligence, teacher tasks and individualized pedagogy (Working Paper). J-PAL.

Fischer, M., Rau, H. A., & Rilke, R. M. (2025). AI Tutoring Enhances Student Learning Without Crowding Out Reading Effort. IZA Discussion Paper, 18338. https://www.iza.org/publications/dp/18338

Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In C. Penstein Rosé et al. (Eds.), *Artificial Intelligence in Education. AIED 2018* (Lecture Notes in Computer Science, Vol. 10947, pp. 154–168). Springer, Cham. https://doi.org/10.1007/978-3-319-93843-1_12

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kim, J. H., Kim, M., Kwak, D. W., & Lee, S. (2021). Home-Tutoring Services Assisted with Technology: Investigating the Role of Artificial Intelligence Using a Randomized Field Experiment. *Journal of Marketing Research, 59*(1), 79-96. https://doi.org/10.1177/00222437211050351

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

Kreijkes, P., Kewenig, V., Kuvalja, M., Lee, M., Hofman, J. M., Vitello, S., Sellen, A., Rintel, S., Goldstein, D. G., Rothschild, D., Tankelevitch, L., & Oates, T. (2026). Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. *Computers & Education, 243*, Article 105514. https://doi.org/10.1016/j.compedu.2025.105514

LearnLM Team, Google, & Eedi (2025). AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. goo.gle/LearnLM-Nov25

Lehmann, M., Cornelius, P. B., & Sting, F. J. (2025). AI meets the classroom: When do large language models harm learning? [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.09047

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence, Vol 6, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Robb, M. B., & Mann, S. (2025). Talk, trust, and trade-offs: How and why teens use AI companions (Research report). San Francisco, CA: Common Sense Media.

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).

SCALE Initiative. (2026). Research Study Repository. https://scale.stanford.edu/ai/repository

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. Computers in Human Behavior, Vol 160, 108386. https://doi.org/10.1016/j.chb.2024.108386

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

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

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

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., Silva, M., & Poulsen, S. (2025). Autograding mathematical induction proofs with natural language processing [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2406.10268

Acknowledgments

Project Coordinator: Imogen Lee

Contributors: Mridul Joshi, Paul Yoo, Ana Trindade Ribeiro, Sofia Wilson, Benjamin Leiva, Cristina Barnard González, Jilli Jung, David Gormley, and Hsiaolin Hsieh

Reviewers: Pilyoung Kim, Carly Robinson, and Isabelle Hau

Images, Design, and Distribution: Julie Brosnan and Claire Fisher Moffet

Funders: Walton Family Foundation and Overdeck Family Foundation

While this review focuses on artificial intelligence, its completion was only possible through the discernment and expertise of our human collaborators. As Project Coordinator, Imogen Lee provided the essential logistical synchronization that kept our work on track.

Our contributors (senior researchers, doctoral students, research associates, and data scientists) performed the rigorous work of validating AI analysis with human judgment to ensure the highest standards of accuracy. The deep expertise of Carly Robinson, Pilyoung Kim, and Isabelle Hau was instrumental in synthesizing these findings into a coherent narrative. We also thank Julie Brosnan and Claire Fisher Moffet for their creative direction and execution in translating complex data into an accessible and engaging resource.

Finally, we extend our sincere appreciation to the Walton Family Foundation, Overdeck Family Foundation, and the Stanford Accelerator for Learning for their foundational support of the AI Hub and the SCALE Initiative.

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 5.2, Gemini 3, and Claude Sonnet 4.6 were used lightly 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.