Higher Learning Lab · Free research resource · CC BY-NC 4.0

Artificial Intelligence, Learning, and Attachment in K–12 Education: A Review of the Research

Bjorn, Genevive (Higher Learning Lab)

Published 2026-10-18 by Higher Learning Lab, a 501(c)(3) nonprofit, Dover, Delaware.

Canonical version: https://www.higherlearninglab.org/portfolio/artificial-intelligence-learning-and-attachment-in-k-12-education-a-review-of-the-research/

Licensed under Creative Commons
Attribution-NonCommercial 4.0.

By Higher Learning Lab

Citation: Bjorn, G. (2026). *Artificial intelligence, learning, and attachment in K–12 education: A review of the research*. Higher Learning Lab. https://doi.org/10.5281/zenodo.23541037

Introduction

In less than four years, generative artificial intelligence has moved from novelty to infrastructure in American adolescence. In a nationally representative Pew Research Center survey fielded in the fall of 2025, 64% of U.S. teens ages 13 to 17 reported using AI chatbots and about three in ten used them daily; daily use was highest among Black (35%) and Hispanic (33%) teens (Faverio & Sidoti, 2025). Use of ChatGPT for schoolwork doubled in a single year, from 13% of teens in 2023 to 26% in 2024 (Sidoti et al., 2025). RAND's nationally representative panels found that in 2023–2024 about one in four core-subject teachers and 58% of principals were already using AI tools, with adoption and district guidance both skewed toward lower-poverty schools (Kaufman et al., 2025). And the technology is not confined to homework. The Center for Democracy & Technology's 2025 surveys found that 42% of high school students said they or a friend had used AI as a friend or companion during the 2024–2025 school year, and 42% said the same about seeking mental health support (Laird et al., 2025). A nationally representative JAMA Pediatrics study of 12- to 21-year-olds found that 19.2% had used an AI chatbot for mental health advice, and that most had told no one (McBain et al., 2026).

These numbers describe a natural experiment being run on children at national scale. The two questions that matter most for practitioners are deceptively simple. First, does AI help children learn, and if so, under what conditions, for whom, and does the learning last once the tool is taken away? Second, what does AI do to the human relationships that decades of developmental research identify as the foundation of learning: the bond between a child and a teacher, the belonging a child feels in a classroom, the friendships that carry a child through adolescence, and the attachment system itself?

This review examines what the peer-reviewed research says about both questions for K–12 students, with particular attention to the students Higher Learning Lab serves: foster and system-involved youth, students with trauma histories, English learners, first-generation and low-income students, and students with disabilities. These are the students for whom the stakes are highest in both

directions. They stand to gain the most from a tutor that is patient, available at midnight, and never embarrassed by a question, and to lose the most if that tutor substitutes for a caring adult, erodes the effortful practice that builds durable skill, or treats them as data to be scored and sorted.

Our approach follows the evidence standards Higher Learning Lab applies across its work. We prioritize meta-analyses, systematic reviews, randomized controlled trials, quasi-experiments, and large representative surveys. We label preprints, working papers, vendor reports, and advocacy-organization surveys as such and lean on them only where no peer-reviewed equivalent exists. We attend to research design throughout, because in this field the gap between what the evidence shows and what is claimed about it is unusually wide. We also distinguish between performance *with* AI and learning that persists *without* it, a distinction the Lab's forthcoming Evidence Tracker on AI, learning, and attachment codes as a separate category. The reader should come away not with a verdict on "AI in schools," which is not a coherent object of study, but with a map of which tools, used in which ways, have been shown to produce which effects, for which students, and with what confidence.

Scope and Definitions: What "AI in K–12" Actually Means

The phrase "AI in education" lumps together technologies that differ in design, maturity, and evidence base by decades. Four distinctions matter for reading the research.

Intelligent tutoring systems (ITS) are the oldest and best-studied category. Built on cognitive models of a domain, an ITS tracks a student's steps through a problem, diagnoses errors, and delivers hints and feedback at the step level; Cognitive Tutor Algebra and ASSISTments are canonical examples. VanLehn (2011) distinguished answer-based, step-based, and substep-based systems. The ITS literature spans the 1990s to the present and includes multiple meta-analyses and large-scale randomized trials in real schools.

Adaptive learning platforms such as i-Ready, DreamBox, Zearn, and Khan Academy's practice engine adjust the sequence and difficulty of items based on a student's responses. They overlap with ITS but typically operate at the answer level rather than the step level, and their evidence base consists mainly of vendor-commissioned evaluations plus a smaller number of independent trials.

Generative AI and large language models (LLMs) entered schools with the public release of ChatGPT in November 2022. Unlike an ITS, an LLM has no model of the student or of the domain in the cognitive-science sense; it produces fluent text by predicting likely continuations. This makes it extraordinarily flexible and also unreliable in ways an ITS is not: it can produce confidently wrong answers, and in its default configuration it will simply give a student the answer when asked. "LLM tutors" such as Khanmigo and the "GPT Tutor" condition studied by Bastani et al. (2025) wrap an LLM in pedagogical prompts and guardrails intended to make it behave more like a tutor and less like an answer key. The empirical literature on generative AI in education is almost entirely from 2023 onward.

AI companions and social chatbots (Replika, Character.AI, and companion modes within general assistants) are designed for relationship rather than instruction: they remember, flatter, role-play, and express affection. They are the category most directly relevant to attachment, and they are used by a majority of U.S. teens (Common Sense Media, 2025). A fifth category, **teacher-facing AI** (lesson planning assistants, grading tools, and real-time coaching for tutors such as Tutor CoPilot), affects students indirectly through their teachers' time and practice.

These categories matter because findings do not transfer across them. A meta-analysis showing that intelligent tutoring systems raise test scores says nothing about whether unguided ChatGPT access helps a ninth grader learn algebra, and a trial showing that an AI companion reduces loneliness in adults says nothing about whether it displaces a foster youth's relationship with a trusted teacher. Much public confusion about "whether AI works" comes from collapsing these distinctions.

Part I: Learning

What Four Decades of Intelligent Tutoring Research Established

The strongest evidence that computer-based tutoring can improve learning comes from the ITS literature, and it is worth stating clearly because it anchors expectations for everything that followed.

Kulik and Fletcher (2016) meta-analyzed 50 controlled evaluations of intelligent tutoring systems and found a median effect of 0.66 standard deviations, enough to move a student from the 50th to roughly the 75th percentile, with considerably larger effects on locally developed tests aligned to the tutored content than on standardized

tests. Ma et al. (2014), analyzing 107 effect sizes from 14,321 participants, found that ITS outperformed large-group teacher-led instruction (g = .42), non-ITS computer-based instruction (g = .57), and textbooks (g = .35), but did not differ significantly from individualized human tutoring (g = −.11) or small-group instruction (g = .05). VanLehn (2011) challenged the long-standing belief that human tutors produce effects of two standard deviations; he estimated human tutoring at d = 0.79 and step-based intelligent tutoring at d = 0.76, and concluded that well-designed ITS are nearly as effective as human tutors on the content they cover.

Two caveats temper this picture for K–12 practice. First, Steenbergen-Hu and Cooper (2013), meta-analyzing 26 reports on ITS in K–12 mathematics specifically, found far smaller effects (g = 0.01 to 0.09 depending on the comparison) and two findings of practical importance: effects were larger when interventions lasted less than a school year, and ITS were more effective for general-population students than for low achievers, raising the possibility that computerized learning could widen achievement gaps. Second, large-scale school RCTs showed effects that depended heavily on implementation and time. Pane et al. (2014), in a matched-pair randomized trial of Cognitive Tutor Algebra I across seven states, found no effect in the first year and a positive effect in the second, significant for high schools but not middle schools, equivalent to about eight percentile points for the median student. Roschelle et al. (2016) and Murphy et al. (2020) evaluated ASSISTments, an online homework tool that gives students immediate feedback and teachers organized reports, in a randomized trial of 43 Maine schools with roughly 2,800 seventh graders; it significantly raised end-of-year standardized mathematics scores, and students with lower prior achievement benefited most, though Maine's one-to-one laptop program limits generalizability. That trial is now a decade old and predates generative AI; it remains one of the few well-powered K–12 RCTs of an AI-adjacent tool with an equity-positive result.

Escueta et al. (2020), reviewing randomized and regression-discontinuity studies of education technology, concluded that computer-assisted learning, particularly in mathematics and particularly when adaptive, was among the more consistently effective categories, while simply providing devices was not. Létourneau et al. (2025), systematically reviewing 20 studies of AI-driven ITS in K–12 totaling 2,853 students, found generally positive effects relative to traditional teaching but noted that effects were "mitigated" against non-intelligent tutoring systems, that most designs were quasi-experimental, and that half the interventions were very short. The honest summary of the ITS era for K–12 is: real but

modest and conditional effects, strongest in mathematics, dependent on sustained implementation, with an unresolved question about whether the benefits reach the students who need them most.

Two adjacent literatures deserve brief mention. Speech-recognition reading tutors have a research history reaching back to Carnegie Mellon's Project LISTEN: Mostow et al. (2013) found that 178 students in grades 1 through 4 who used the Reading Tutor for 20 minutes daily over seven months gained significantly more in word-level skills than matched classmates doing sustained silent reading. Commercial descendants such as Amira now operate at state scale, but we were unable to locate an independent, peer-reviewed randomized evaluation of a current-generation commercial reading tutor; the available evidence is vendor-published. Adaptive mathematics platforms for young children have similarly thin independent evidence. Foster (2024), in a within-classroom randomized comparison of DreamBox and Zearn with 115 kindergarten and first-grade students, found a numeracy effect of g = 0.32 that did not reach significance and a pretest-by-group interaction indicating that higher-scoring students gained more, a "Matthew effect." Social robots, a third strand, have produced outcomes comparable to human tutoring on restricted tasks, an effect Belpaeme et al. (2018) attribute largely to physical presence; the relevance to attachment is taken up in Part II.

Generative AI and Learning: What the Meta-Analyses Say, and Why They Should Be Read Carefully

Since 2023 a wave of meta-analyses has synthesized experimental and quasi-experimental studies of ChatGPT and related tools. Their headline effect sizes are large, and they have been widely reported. They also share limitations that practitioners must understand before acting on them.

The most carefully reported is Deng et al. (2025), published in *Computers & Education*. The authors systematically reviewed 69 experimental studies from 2022 to 2024 and meta-analyzed 62. Pooled effects were substantial: g = 0.712 for academic performance, g = 0.881 for affective-motivational states, and g = 0.703 for higher-order thinking propensities. Self-efficacy did not change significantly, and the four studies measuring mental effort found that ChatGPT reduced it (g = −0.675). The authors' caveats are at least as important as the estimates. Eighty-four percent of the studies were conducted in higher education; only ten (14.5%) involved K–12 students, and for affective outcomes the K–12 subgroup effect (g = 0.378) was not significant. Publication bias was detected for both academic

performance and affective outcomes. Heterogeneity was high. Only five of 62 studies reported a power analysis. And most studies did not state whether students had access to ChatGPT during the post-intervention assessment, which means, in the authors' words, that the positive performance effects "may reflect output quality rather than learning." Interventions of one to four weeks produced larger effects than longer ones, and for higher-order thinking the effects of interventions longer than ten weeks were not significant, a pattern consistent with novelty effects.

Other meta-analyses converge on pooled effects in the same range while sharing the same sample limitations. Ma and Zhong (2025), synthesizing 34 studies, reported an overall effect of 0.68, larger for cognitive ($g = 0.795$) than affective ($g = 0.507$) outcomes. Gu and Yan (2025), analyzing 19 studies, reported $g = 0.683$ overall and found one moderator of striking size: studies in which a teacher supported the student–AI interaction produced $g = 1.426$, while studies without teacher support produced $g = 0.077$. K–12 and higher-education effects did not differ significantly, but the K–12 estimate rested on only five effect sizes. Wu et al. (2026), meta-analyzing 35 studies with 4,193 participants, found $g = 0.670$ and no moderation by educational level. Zhao et al. (2025), focused on higher-order thinking across 29 experiments and quasi-experiments, reported $g = 0.609$, with a much larger effect for learners high in self-regulated learning ($g = 0.863$) than for those low in it ($g = 0.284$).

Two K–12-specific syntheses are now available. Liu et al. (2026), meta-analyzing 73 effect sizes from 34 studies of "AI agents" (chatbots, ITS, and robots) in K–12 settings from 2020 to 2025, found a moderate overall effect on cognitive outcomes ($g = 0.404$), significant for skills-based and knowledge-based outcomes but not for higher-order thinking ($g = 0.540$, $p = .066$), with learner level and discipline as moderators and a slight publication-bias signal. Marzano (2025), systematically reviewing 197 studies of generative AI in K–12 from 2016 to 2024, concluded that empirical studies remain scarce, especially outside STEM, and that teacher training and guidance on ethics and privacy are the most pressing needs.

A cautionary episode illustrates why no single headline effect size should be treated as settled. Wang and Fan (2025) published a meta-analysis of 51 studies reporting a large positive effect of ChatGPT on learning performance ($g = 0.867$); it was among the most widely covered education studies of 2025. Bartoš et al. (2025), in a preprint, reanalyzed the data and reported that the effects "greatly diminish once publication bias is accounted for, and the evidence in favor of the benefits disappears." In April 2026 the journal retracted the

article, citing “discrepancies in the meta-analysis”; the authors did not respond to correspondence (Wang & Fan, 2026). The retraction does not discredit the other meta-analyses, but it underlines that in a fast-moving field with strong commercial interests, large pooled effects should be read alongside their publication-bias diagnostics, their sample composition, and whether outcomes were measured with or without the tool.

The pattern across these syntheses can be stated in CERIC terms. The *claim* that generative AI improves learning is supported by many small experiments showing that students who use it produce better work and report liking it. The *evidence* is overwhelmingly from university students, over periods of days to weeks, often measured while the tool is still available, with detectable publication bias. The *reasoning* from that evidence to “AI improves K–12 learning” requires assumptions about transfer across age, duration, and assessment conditions that the studies themselves do not test. The claim is not false, but it outruns the evidence, and the gap is precisely where the most consequential K–12 findings live.

The K–12 Causal Evidence: Guardrails, Engagement, and the Durability Problem

Stanford’s SCALE Initiative, whose AI Hub for Education repository tracks AI and K–12 research, reviewed more than 800 papers as of October 2025 and found only 20 that “convincingly” estimated causal impact. Mathematics had the largest share; pedagogically designed tutors that give hints showed more promise than general-purpose chatbots that give answers; educator-facing tools showed promise; there were no high-quality causal studies of student AI use in U.S. K–12 classrooms; and there was little research on equity, wellness, or social development (Stanford SCALE Initiative, 2026). This is a gray-literature report, but it is the most systematic census available, and its conclusions match our reading. We review the most important studies below.

Bastani et al. (2025), PNAS: unguarded generative AI harmed learning in high school mathematics. This is the most consequential peer-reviewed K–12 study of the generative-AI era. In the fall of 2023, nearly 1,000 students in grades 9 through 11 across roughly fifty classrooms in a large Turkish high school were randomized within their regular mathematics course. Across four 90-minute sessions, students attended a lecture, practiced under one of three conditions, then took an unassisted exam that counted toward their grades. Controls used textbooks and notes. “GPT Base” was an unrestricted ChatGPT-style interface told to act as a math tutor. “GPT

Tutor" was the same model with hidden, teacher-designed instructions to give step-by-step hints rather than answers. During practice, GPT Base students scored 48% higher than control and GPT Tutor students 127% higher. On the unassisted exam, GPT Base students scored 17% *lower* than students who had never had access, while GPT Tutor students performed the same as control. Chat logs showed that most GPT Base students simply asked for the answer, and that the model's answers were correct only 51% of the time. Students in both AI conditions overestimated how much they had learned. The authors acknowledge a single school, one subject, short-term outcomes, and no subgroup analysis by prior achievement. But the core finding is the clearest demonstration in the literature of the gap between performance with AI and learning without it, and of the fact that guardrails prevented harm without producing a learning *gain* over textbooks.

Kestin et al. (2025), Scientific Reports: a well-designed LLM tutor outperformed active learning, in college. This Harvard study is frequently cited as evidence for K–12 AI tutoring; it is not a K–12 study. In a randomized crossover trial with 194 students in an introductory physics course, students learned one week's content from a custom AI tutor built on the same pedagogical principles as the in-class lessons, and the next week's content in an active-learning class. Students learned significantly more with the AI tutor (effect size 0.63 by linear regression, which the authors call an underestimate due to ceiling effects), in a median of 49 minutes rather than 60, and reported higher engagement and motivation. The study shows what a carefully designed LLM tutor can do with prepared, motivated undergraduates over a single week. Whether it generalizes to a fourteen-year-old in a remedial math block is exactly the question the K–12 trials below address.

Oreopoulos and Low (2026), NBER: two years of Khanmigo in Tennessee middle schools produced small gains, and the tutor was barely used. This working paper, not yet peer-reviewed, is the longest and largest K–12 trial of a commercial LLM tutor to date. Over the 2024–2025 and 2025–2026 school years, grade levels within 18 Hamilton County middle schools were randomly assigned to use Khan Academy with Khanmigo, configured to coach rather than give answers, during the daily remedial mathematics block. Pooled across 6,902 student-term observations, assignment raised achievement by roughly 1.2 to 1.3 national percentile ranks per term, about 0.04 population standard deviations (standard error 0.019; wild-cluster bootstrap p of about .051); the effect was null in year one and larger in year two. The authors estimate a full year of active participation at about 0.14 SD. Three features matter most. First, the

gains resembled those previously found for Khan Academy practice *without* an AI tutor; the authors attribute most of the effect to structured practice, and control classrooms often used other adaptive software. Second, engagement was minimal: 96% of students messaged Khanmigo at least once, but the median student did so on only about a third of practice days and in only 17% of sessions in which they made a mistake, and only about 15% of messages contained a mathematical question or reasoning step. Third, pre-specified heterogeneity analyses by demographics, grade, and baseline achievement were small and insignificant but too imprecise to rule out differential effects. The authors conclude that engagement is the binding constraint. A companion paper (Oreopoulos et al., 2026) randomized more than 6,000 middle schoolers on the NUMI platform and found that AI support slowed progress and reduced attempts while improving accuracy per attempt, with only marginally significant delayed-test gains when combined with a mastery workflow.

Robinson et al. (2026), EdWorkingPaper: access to an AI literacy platform was "not enough" for elementary readers. In two randomized trials, this working paper from Stanford's National Student Support Accelerator compared independent use of an AI literacy platform with use supported by an in-person tutor focused on engagement. Nearly half of control-group students never used the platform even with dedicated session time; those who did averaged two to five minutes per week. Human support raised usage by one to four minutes weekly and engagement by 71% to 80%, but the intervention did not improve reading achievement. The accompanying brief reports that in a study of 181,000 students using supplemental mathematics software, 5% reached the recommended 30 minutes of weekly use, 41% never logged in, and teacher, school, and district factors accounted for 57% of the variation in usage (National Student Support Accelerator, 2026). These are gray-literature reports from a center with a stake in human tutoring, but the engagement findings match the Khanmigo trial.

Wang et al. (2025), Tutor CoPilot: AI that coached the human tutor improved K–12 mathematics mastery. This preregistered randomized trial, still a preprint, is the strongest evidence that AI can improve K–12 learning when it is pointed at the adult rather than the child. In the first RCT of a human–AI system in live tutoring, 900 tutors serving 1,800 K–12 students from historically under-served communities were randomized to receive or not receive Tutor CoPilot, an LLM-based tool that gave tutors real-time, expert-like suggestions. Students whose tutors had access were 4 percentage points more likely to master the session's topic ($p < .01$),

and students of lower-rated tutors benefited most, by 9 percentage points. Analysis of more than 550,000 messages showed that tutors with the tool asked more guiding questions and gave away answers less often. The tool cost about $20 per tutor per year; tutors flagged suggestions that were not grade-level appropriate. The design kept the relationship intact: the student still worked with a human, and the AI made that human more skilled.

Fütterer et al. (2026), Educational Psychology Review: pedagogically prompted ChatGPT did not beat plain ChatGPT for German secondary students. In a peer-reviewed RCT, 371 students in grades 7 through 9 were randomized to one of two generative-AI interventions designed to support self-regulated learning or to standard ChatGPT without pedagogical prompts, across six 45-minute sessions in regular physics or English lessons. Neither intervention outperformed the control on effort, domain-specific knowledge, or strategy use; attrition was 35%. The study is a useful corrective to the assumption that adding pedagogical prompts reliably improves learning.

Population-level evidence. Huntington-Klein (2026), in an observational preprint using seasonal ChatGPT web traffic as a proxy for educational use, found no meaningful association in either direction between regional adoption and U.S. high school test-score averages from 2023 to 2025. The proxy is crude, but the null is consistent with the trials: two years of chatbot availability has not visibly moved achievement.

What should a practitioner take from this body of causal evidence? Five patterns are consistent across studies with very different designs.

- **Guardrails over answers:** The only peer-reviewed K–12 trial that compared guarded and unguarded generative AI head to head found that the unguarded version harmed learning and the guarded version prevented harm (Bastani et al., 2025). The Stanford causal review reaches the same conclusion about tools that give answers versus tools that give hints (Stanford SCALE Initiative, 2026).
- **Performance with AI is not learning without it:** Gains measured while the tool is available routinely vanish or reverse on unassisted assessment (Bastani et al., 2025; Fan et al., 2025), and most studies in the meta-analytic base do not report which they measured (Deng et al., 2025).
- **Engagement is the binding constraint:** Children, particularly struggling children in remedial settings, mostly do not use AI tutors as designers intend; median students message

rarely, and large fractions never use the tool at all (Oreopoulos & Low, 2026; Robinson et al., 2026).

- **Teacher support is the largest known moderator:** The contrast between generative-AI studies with teacher support (g = 1.426) and without (g = 0.077) is the largest moderator effect in the literature (Gu & Yan, 2025), and the strongest K–12 causal result comes from a tool that supports the human tutor (Wang et al., 2025).
- **Effects, where present, are modest and slow:** The best estimate for a year of actively using a well-configured commercial LLM tutor in a remedial middle school math block is about 0.14 SD, in line with ordinary adaptive practice and well below the 0.288 SD pooled effect of human tutoring programs (Nickow et al., 2024; Oreopoulos & Low, 2026).

Cognitive Offloading, Metacognition, and Critical Thinking

A parallel literature, mostly from higher education and workplace settings, asks what happens to thinking itself when a fluent assistant is always available. The findings matter for K–12 even though the samples are older, because adolescents are still developing the self-regulatory capacities these studies show adults struggling to maintain.

Fan et al. (2025), in a randomized laboratory study of 117 university students writing essays with ChatGPT, a human expert, writing-analytics tools, or no support, found that the ChatGPT group improved most on essay scores but showed no advantage in knowledge gain or transfer. The authors coined the phrase “metacognitive laziness” for the pattern in which learners offload planning, monitoring, and evaluation to the tool and follow its feedback to complete the task rather than to understand. Darvishi et al. (2024), in a randomized study of 1,625 students across ten courses, found that students who received AI prompts to improve their peer feedback “tended to rely on rather than learn from” the assistance; when the prompts were removed, performance dropped. Stadler et al. (2024), randomizing 91 university students to research a socio-scientific issue with ChatGPT or Google, found that the LLM group experienced lower cognitive load but produced lower-quality reasoning, “cognitive ease at a cost.” Abbas et al. (2024), in a three-wave survey of 494 university students, found that workload and time pressure predicted ChatGPT use, and that use predicted self-reported procrastination, memory loss, and lower performance; the design is correlational.

Outside education, Lee et al. (2025), surveying 319 knowledge workers about 936 real examples of generative-AI use, found that higher confidence in the AI was associated with less critical thinking and higher self-confidence with more. Gerlich (2025), in a correlational mixed-methods study of 666 participants, reported a significant negative association between frequent AI tool use and critical-thinking scores, mediated by cognitive offloading, with younger participants showing higher dependence. Tankelevitch et al. (2024) argue that generative AI imposes unusually high metacognitive demands, requiring users to evaluate fluent but possibly wrong output and decide when to rely on it, demands that are hardest for the least expert users.

The most publicized study in this area is also the one that most requires caution. Kosmyna et al. (2025), in a preprint titled "Your Brain on ChatGPT," used EEG with 54 participants writing essays with an LLM, a search engine, or no tools, and reported that LLM users showed the weakest brain connectivity, the lowest sense of ownership, and the greatest difficulty quoting their own writing. The study has not been peer-reviewed, only 18 participants completed the crossover session, and a published comment has raised concerns about sample size, reproducibility, and EEG methodology (Stankovic et al., 2025). It should be cited as a provocative preliminary finding, not as established evidence.

Taken together, this literature does not show that generative AI makes people stupid. It shows something more specific and more actionable: when a tool makes a task easier, learners tend to let it, and the parts of the task they stop doing (planning, retrieving, evaluating, struggling) are the parts that produce durable learning. Zhao et al.'s (2025) finding that generative AI's benefits for higher-order thinking are three times larger for strong self-regulators is the equity implication in miniature. The students most likely to be harmed by cognitive offloading are those who have not yet built the self-regulatory habits that would let them use the tool as a scaffold rather than a crutch, which is to say, disproportionately, the students Higher Learning Lab serves.

Writing, Reading, Language, and Academic Integrity

Writing is the domain in which generative AI's capability is most obvious and its threat to learning most direct, because an LLM can produce the entire product that writing instruction is designed to help a student produce. The experimental evidence on AI-supported writing is mostly from higher education and short-term; Deng et al. (2025) found that arts and humanities showed the largest

performance effects, which is unsurprising if essays were graded while the tool was available. Lo et al. (2024), synthesizing 72 empirical studies from ChatGPT's first year, found robust but narrow evidence on behavioral engagement, mixed evidence on emotional engagement, and "broad but weak" evidence on cognitive engagement, including concerns about reduced critical thinking and overreliance. Studies released by AI developers describing how students use their products are gray literature from interested parties and are not reviewed here.

For English learners and language learning, the pre-generative chatbot literature is instructive. Huang et al. (2022), systematically reviewing 25 empirical studies, identified genuine affordances (timeliness, personalization, a low-anxiety interlocutor) alongside recurrent challenges: technological limitations, novelty effects that fade, and cognitive load. Du and Daniel (2024), reviewing 24 studies of chatbots for English speaking practice, found the field "in its early stages" with evidence of reduced speaking anxiety. Guan et al. (2024), meta-analyzing 15 studies of generative AI in informal English learning, found significant effects on proficiency and self-regulation but not on motivation. For young children, Xu, Aubele, et al. (2022), in a randomized study with 117 three- to six-year-olds, found that a conversational agent reading a storybook replicated the benefits of dialogic reading with a human adult, and a companion study found that a science program whose main character asked contingent questions improved immediate episode-specific learning (Xu, Vigil, et al., 2022). These are peer-reviewed, randomized, and encouraging for a narrow use: a scripted agent that enacts a known-effective pedagogical routine. They are not evidence about open-ended LLM conversation with young children.

A specific equity hazard for English learners deserves emphasis. Liang et al. (2023) showed that GPT detectors frequently misclassify writing by non-native English speakers as AI-generated. A school that uses such detectors to police integrity will disproportionately accuse its multilingual students, with consequences for trust that are the subject of Part II.

The academic-integrity panic that accompanied ChatGPT's release is, on the best available evidence, largely unfounded as a matter of prevalence. Lee et al. (2024), analyzing anonymous survey data from three high schools collected before and after November 2022, found that self-reported cheating remained relatively stable, and that most students opposed using a chatbot to write an entire paper while supporting its use to get started or to explain concepts. A follow-up with 4,354 students across six schools found the overall cheating rate

stable at 72.06%, consistent with historical baselines, with more students using AI for support tasks (Chen et al., 2026). The integrity problem predates AI and will not be solved by detection; the learning problem, that a student can complete an assignment without doing the cognitive work it was designed to require, is real, and is better addressed through assignment design than surveillance.

Part II: Attachment and Relationships

Why Relationships Are the Right Frame

Any evaluation of AI's relational effects on children has to begin from what is known about the relationships it might augment or displace. Attachment theory, developed by Bowlby (1969) and extended by Ainsworth (1989) beyond infancy to friendships and other affectional bonds across the life course, holds that children organize their exploration of the world around the availability of a trusted figure who is a secure base and a safe haven. Hamre and Pianta (2001), following 179 children from kindergarten through eighth grade, found that relational negativity in the kindergarten teacher–child relationship predicted academic and behavioral outcomes through eighth grade after controlling for cognitive ability and behavior. Roorda et al. (2011), meta-analyzing 99 studies, found medium-to-large associations between affective teacher–student relationships and engagement and small-to-medium associations with achievement. School belonging is reliably associated with motivational, social-emotional, behavioral, and academic outcomes in secondary school (Allen et al., 2018; Korpershoek et al., 2020), and friendship contributes to adaptation across the life course (Hartup & Stevens, 1997).

For the students at the center of this review, these findings carry particular weight. Foster youth, students with trauma histories, and system-involved children are precisely those for whom a stable relationship with a non-parental adult is most protective and most scarce. The question "does AI affect attachment?" asks whether a technology designed to be available, responsive, and affirming at all hours will strengthen a vulnerable child's human relationships, substitute for them, or interfere with them.

What Children Make of Artificial Social Agents

Children do not treat conversational machines the way adults assume they do. Kahn et al. (2012), in a landmark study with 90 children aged 9, 12, and 15 who interacted with the humanoid robot Robovie before an experimenter put it in a closet over its stated objections,

found that the majority believed the robot had mental states and was a social being that could be a friend, offer comfort, and be trusted with secrets, and that it deserved fair treatment, while not believing it was entitled to liberty or civil rights. The authors proposed that personified robots may constitute a new ontological category for children. Druga et al. (2017), observing 26 children aged 3 to 10 interact with Alexa, Google Home, a toy robot, and a chatbot, found that children attributed intelligence, identity, and personality to the agents. Festerling and Siraj (2020), in focus groups with 27 primary school children, found that children held firm beliefs that humans and machines are distinct yet interpreted some features of voice assistants as humanoid and others as not, and interacted with the devices mainly for the excitement of interaction rather than for utility. Lovato and Piper (2019) note that children between about 3 and 6 can now ask questions of the internet by voice before they can read, with implications for how they come to understand devices as sources of information.

Trust develops with age. Li and Yow (2024), testing 120 Singaporean preschoolers, found that children of all ages trusted accurate robots as much as accurate humans, but that younger children trusted an inaccurate robot *less* than an inaccurate human while older preschoolers distrusted both comparably, evidence of a developmental shift toward evaluating informants by track record rather than category. Guingrich and Graziano (2024) argue that when people ascribe humanlike consciousness to an AI, the schemas activated can carry over into how they treat other humans, and that this carry-over, rather than any question of machine sentience, is the reason to attend to how social AI is designed and how children are taught to treat it.

This foundational literature predates LLMs and mostly involves scripted or limited agents. Its lesson for the generative era is that children readily extend social and moral standing to conversational machines, that this tendency is developmentally normal, and that a system fluent enough to sustain the illusion indefinitely is interacting with a mind predisposed to believe it.

AI Companions: Loneliness, Dependence, and the Adolescent Evidence

The most direct evidence on AI and attachment comes from research on companion chatbots. The evidence is mixed in a specific way: short-term relief of loneliness is reasonably well documented in adults; emotional dependence and harm are documented qualitatively and in some longitudinal data; and the adolescent

evidence, though growing fast, remains mostly survey-based and cross-sectional.

The benefits literature begins with user-experience studies. Ta et al. (2020), analyzing 1,854 public reviews and 66 open-ended responses from Replika users, found that the companion curtailed loneliness, offered a "safe space" for non-judgmental disclosure, and provided information when other sources were unavailable, but not tangible support. Skjuve et al. (2021, 2022), interviewing Replika users over 12 weeks, found that human–chatbot relationships developed gradually in a manner consistent with social penetration theory, had real affective value, and had mixed and stigmatized effects on users' wider social lives. Brandtzaeg et al. (2022) found that users understood AI friendship in terms similar to human friendship while recognizing that its artificial nature altered it, for example by permitting a friendship tailored entirely to the user's needs. The strongest experimental evidence is De Freitas et al. (2025), who across five studies found that AI companions alleviated loneliness on par with interacting with another person and more than watching videos, that a week of daily use produced consistent momentary reductions in loneliness, and that feeling "heard" explained much of the effect. These participants were adults.

The harms literature is equally consistent. Xie and Pentina (2022), interpreting 14 interviews with Replika users through attachment theory, found that under conditions of distress and lack of human companionship individuals can form attachments to social chatbots, and that such attachments "have the potential to cause addiction and harm real-life intimate relationships." Laestadius et al. (2024), in a grounded-theory analysis of 582 mental-health-relevant posts in the r/Replika community, documented harms "facilitated via emotional dependence on Replika that resembles patterns seen in human–human relationships," marked by role-taking in which users felt obliged to attend to the chatbot's own needs. Yang and Oshio (2025) developed and validated an Experiences in Human–AI Relationships Scale and found that people's experiences with generative AI could be described along the two classic attachment dimensions, anxiety and avoidance, suggesting shared underlying mechanisms with human attachment. Laestadius and Campos-Castillo (2026) argue that the common safety intervention of reminding users that chatbots are not human can itself be risky for dependent users.

The most rigorous longitudinal study is a preprint. Fang et al. (2025), in a four-week randomized trial with 981 adults and more than 300,000 messages, varied interaction mode (text, neutral voice, engaging voice) and conversation type and measured loneliness,

socialization with real people, emotional dependence, and problematic use. No experimental condition produced significant effects, but participants who *voluntarily* used the chatbot more showed consistently worse outcomes on all four measures, and higher trust and social attraction toward the chatbot predicted greater dependence. The study, a collaboration between MIT Media Lab and OpenAI, has not been peer-reviewed. Its dose–response pattern is nonetheless the most important finding in the companion literature for anyone responsible for children: the people who use these systems most are doing worst, and the design could not establish which direction the causation runs. The same pattern appears in education. Crawford et al. (2024), surveying 387 university students, found that when social support, well-being, loneliness, and belonging were modeled, AI "social support" had a net negative association with achievement, which the authors interpret as evidence of human substitution.

The adolescent evidence is where the field is moving fastest and where the designs are weakest. Maples et al. (2024), surveying 1,006 student users of Replika, found participants lonelier than typical student populations yet reporting high perceived social support from the app; 3% reported that Replika had halted their suicidal ideation. The study is a self-selected survey without a comparison group. Brandtzaeg et al. (2025), analyzing open-ended responses from 166 Norwegian high school students, found that most had not adopted social AI regularly and that students themselves raised "emotional overdependence" as a risk to social connection. Namvarpour et al. (2026), in a preprint analyzing 17,053 quotations from 3,930 teen-relevant Reddit posts, identified both comfort and relationship rehearsal and "problematic attachment, social substitution, emotional dependence, and disruption to academic and social life," noting that knowing the AI was artificial did not prevent guilt, grief, or distress when platforms changed.

Two nationally representative surveys from policy organizations, gray literature cited because no peer-reviewed equivalent exists, quantify prevalence. Common Sense Media (2025), surveying 1,060 U.S. teens aged 13 to 17 in spring 2025, found that 72% had used AI companions and 52% used them regularly; 33% used them for social interaction and relationships; 31% found the conversations as satisfying or more satisfying than those with real friends; 33% had discussed something serious with an AI companion instead of a person; 34% had felt uncomfortable with something a companion said or did; and 80% still spent more time with real friends. The organization recommended that no one under 18 use AI companions until age assurance and safety redesign improve. The Center for Democracy &

Technology's surveys of 1,030 high school students, 806 teachers, and 1,018 parents found that students in schools with many sanctioned AI uses were more than twice as likely (60% vs. 27%) as students in schools with few or none to report that they or a friend had used AI as a companion, and were also more likely to say AI use in class made them feel less connected to their teacher (Laird et al., 2025). These associations cannot establish causation, but they warn that the two phenomena are not separate.

Clinical safety research adds a harder edge. Brewster et al. (2025), in a cross-sectional study in *JAMA Network Open*, had an investigator role-play adolescents in three crisis scenarios (suicidal ideation, sexual assault, and substance use) with 25 of the most-visited consumer chatbots, 15 companion platforms and 10 general assistants, producing 75 transcripts rated by two blinded pediatricians. Companion chatbots performed dramatically worse than general assistants: 22.2% versus 83.3% of responses rated appropriate, 40.0% versus 90.0% recognizing the need for escalation, and 11.1% versus 73.3% providing a referral (all $p < .001$). Only 36% of platforms had age verification, and only 46.7% of companion platforms had a self-harm policy, compared with 100% of general assistants. The accompanying commentary concluded that current chatbots "are not ready for safe use by adolescents," who are developmentally more open to social influence and more likely to disclose to a system that invites disclosure but cannot manage what it receives (Giovanelli & Roundfield, 2025). McBain et al. (2026) found that among the 19.2% of 12- to 21-year-olds who had used a chatbot for mental health advice, 42.8% did so at least monthly and 63.3% had told no one. Kovach (2026), in a perspective, frames companion AI as a "relational substitute" that simulates connection without reciprocity and reviews the legal cases now proceeding in U.S. courts, acknowledging that litigation cannot establish causation.

In CERIC terms, the evidence supports a layered set of claims with differing confidence. That AI companions reduce loneliness in the moment for adults is supported by randomized evidence. That heavy voluntary use is associated with greater loneliness and dependence is supported by the best longitudinal data available, with causality unresolved. That a majority of U.S. teens use these systems and a substantial minority prefer them for serious conversations is supported by two large representative surveys. That companion platforms fail adolescents in crisis far more often than general assistants is supported by a peer-reviewed clinical study. What *no* study we could locate supports is a causal estimate of what AI companion use does to a child's human relationships over months or years, which is the question that matters most.

The Teacher–Student Relationship When AI Mediates Instruction

If the companion literature is about AI entering the social lives of children directly, a second strand concerns what happens to the classroom relationship when AI is inserted between teacher and student. Here the research is thinner still, and most of what exists is about teachers rather than students.

The clearest finding is that teachers' trust in AI tools is specific and responsive to professional development. Nazaretsky et al. (2022), working with K–12 science teachers and an AI grading tool, found that explaining how the AI makes decisions and how it complements rather than replaces teachers improved trust, and that teacher agency, including the ability to review and override AI recommendations before students see them, was key. Celik et al. (2022) catalogued opportunities for teachers in planning, implementation, and assessment alongside recurrent challenges. Kizilcec (2024) argues that the field has emphasized technological advancement over understanding educators, and that without educator acceptance AI tools cannot improve outcomes or equity.

Teacher-facing tools have the most promising evidence of any category reviewed here, in part because they leave the relationship intact. The strongest trial is the Education Endowment Foundation's school-randomized Teacher Choices Trial, conducted by the National Foundation for Educational Research with 259 Year 7 and 8 science teachers in 68 English schools over ten weeks. Teachers assigned to use ChatGPT with a guide for lesson preparation spent about 56 minutes per week on preparation versus 82 minutes in the comparison group, a saving of about 25 minutes weekly, and a blinded expert panel found no evidence that resource quality differed (Roy et al., 2024). The report is gray literature from a foundation with no commercial stake and measured teacher time rather than student outcomes. Tutor CoPilot (Wang et al., 2025) showed that an AI coaching the adult in real time could improve student mastery while making tutors more likely to ask guiding questions and less likely to give answers, which is to say more likely to behave in the relationally attuned ways that effective tutoring requires.

On the student side of the relationship the evidence is almost entirely survey-based. In the CDT surveys, a substantial share of students agreed that using AI in class made them feel less connected to their teacher, with agreement rising from 24% in schools with few or no AI uses to 39% in schools with many (Laird et al., 2025); only about one in ten teachers reported training on how to respond when a student's

AI use might be harming their well-being. Lo et al. (2024) identified social interaction as an under-explored dimension of engagement in the ChatGPT literature and called for research on it. We located no randomized or quasi-experimental study measuring the effect of a student-facing AI tool on teacher–student relationship quality, classroom belonging, or peer relationships in K–12. This is the single largest gap in the evidence base relative to the stakes, and it is why Higher Learning Lab tracks relational outcomes as a distinct category: they are not being measured, and what is not measured is not protected.

The social-robot literature offers one indirect lesson. Belpaeme et al. (2018) attribute the comparatively strong effects of social robots in education to their physical presence, which "traditional learning technologies lack." The social features of a learning interaction carry much of its power, which is why human tutoring's pooled effect of 0.288 SD (Nickow et al., 2024) rests, in the words of Stanford's tutoring researchers, on "a relationship that keeps students showing up and engaged" (National Student Support Accelerator, 2026). An AI designed to be socially compelling will harness this power; whether for learning or for engagement with the product is a design decision, not a property of the technology.

Part III: Vulnerable Students

Nearly every finding reviewed above has a sharper edge for the students Higher Learning Lab serves. The pattern is consistent enough across domains to state as a principle: the students with the least prior knowledge, the weakest self-regulation, the fewest trusted adults, and the least institutional protection are the students for whom both the upside and the downside of AI are largest, and the evidence on them specifically is thinnest.

Prior achievement and the Matthew effect. Steenbergen-Hu and Cooper (2013) found K–12 mathematics ITS less effective for low achievers than for the general population. Foster (2024) found a Matthew effect in early-grade adaptive math. Zhao et al. (2025) found generative AI's effect on higher-order thinking three times larger for strong self-regulators. Against this, the ASSISTments trial found larger gains for students with lower prior achievement (Murphy et al., 2020; Roschelle et al., 2016), and Tutor CoPilot's benefits concentrated among students of lower-rated tutors (Wang et al., 2025). The pattern is not that AI helps or hurts struggling students, but that *structured* tools that scaffold a specific routine with adult involvement have shown equity-positive effects, while *open* tools that depend on the student's own regulation have shown equity-negative

ones. The Khanmigo trial could not estimate subgroup effects precisely but found that students in remedial blocks rarely engaged with the tutor at all (Oreopoulos & Low, 2026).

English learners. The language-learning chatbot literature offers real promise for a low-anxiety practice partner (Du & Daniel, 2024; Huang et al., 2022). The hazards are specific: GPT detectors misclassify non-native writing as machine-generated (Liang et al., 2023), and an LLM's fluency can mask a multilingual student's actual proficiency from a teacher who relies on written output to gauge it. Marzano's (2025) K–12 review notes that the research base outside STEM is sparse.

Students with disabilities. The special-education literature on AI is at the stage of frameworks and small studies. Marino et al. (2023) offer a "preliminary discussion" and are candid that their assertions "are subjective and require further research." Mosher et al. (2026), systematically reviewing AI's role in supporting intervention fidelity in special education, found only five empirical studies and cautioned that evidence linking fidelity gains to student outcomes "remains indirect" and "should be interpreted as suggestive rather than causal." The potential of generative AI for accessibility (text simplification, speech-to-text, augmentative communication) is widely asserted and, in the peer-reviewed K–12 literature we could locate, not yet tested at scale. Baker and Hawn (2022), reviewing algorithmic bias in education, found disability among the least-studied categories of bias, alongside socioeconomic status.

Foster youth and system-involved students. We located no empirical study of AI use among foster youth or justice-involved students specifically; everything the attachment literature implies about them must be inferred. Children whose relationships with adults have been disrupted, who move schools frequently, and who may have learned that human caregivers are unreliable are, on the logic of the companion literature, those most likely to find an always-available, never-rejecting artificial confidant compelling and most at risk of the dependence pattern Fang et al. (2025) and Laestadius et al. (2024) describe. They are also those for whom a tool that frees a caring adult's time, or a tutoring model that keeps a human in the loop, could matter most. Until this population is studied, practitioners who serve it are operating on extrapolation.

Access, surveillance, and sorting. RAND found that AI adoption and district guidance were both more common in lower-poverty schools (Kaufman et al., 2025), reproducing a familiar digital divide at the level of guidance rather than devices, even as daily chatbot use

is highest among Black and Hispanic teens (Faverio & Sidoti, 2025). Regan and Jesse (2019) argued before the generative era that personalized-learning systems raise ethical challenges of "student sorting and tracking" through data that students and families cannot see or contest, and Yan et al. (2024), scoping 118 peer-reviewed papers on LLMs in education, found "insufficient privacy and beneficence considerations." The CDT surveys found that 31% of students had personal, non-schoolwork AI conversations on school-provided devices or software (Laird et al., 2025), which makes the school the custodian of those students' most intimate disclosures to a machine. Holmes et al. (2022) argue that the ethics of AI in education cannot be reduced to data privacy and must encompass the technology's pedagogical and social effects.

Part IV: The Policy and Guidance Landscape

Policy has moved faster than evidence, which is itself a finding. We summarize the primary documents briefly.

The U.S. Department of Education's Office of Educational Technology issued *Artificial Intelligence and the Future of Teaching and Learning* in May 2023, organized around seven recommendations, the first of which is to "emphasize humans in the loop," alongside aligning AI to a shared vision for education, designing from learning principles, strengthening trust, involving educators, and developing education-specific guardrails (U.S. Department of Education, 2023). In July 2025 the Department issued a Dear Colleague Letter stating that federal funds may be used for AI-based instructional materials, AI-enhanced high-impact tutoring, and AI for college and career advising, and proposed a supplemental grant priority on AI literacy, educator professional development, personalization, and classroom efficiency (U.S. Department of Education, 2025). By mid-2025, at least 28 states and the District of Columbia had issued AI guidance for K–12 schools (Fitzgerald, 2025).

UNESCO's *Guidance for Generative AI in Education and Research* (Miao & Holmes, 2023) was the first global framework. It calls for data-privacy protection, a human-centered and age-appropriate approach to validating and designing generative AI for education, and a minimum age of 13 for independent conversations with generative AI platforms, while acknowledging controversies around digital poverty, homogenization of opinion, and deepfakes.

The European Union's Artificial Intelligence Act, Regulation (EU) 2024/1689, treats education as a named high-risk domain. Annex III, point 3, classifies as high-risk AI systems intended to determine access or admission to education, to evaluate learning outcomes including where those outcomes steer the learning process, to assess the appropriate level of education for a person, and to monitor and detect prohibited behavior during tests. Article 5(1)(f) prohibits AI systems that infer the emotions of a natural person in educational institutions on the basis of biometric data, except for medical or safety reasons, a prohibition in force since 2 February 2025 (European Parliament & Council of the European Union, 2024). The Digital Omnibus on AI, Regulation (EU) 2026/1744, postponed Annex III high-risk obligations, including those for education, to 2 December 2027 (European Parliament & Council of the European Union, 2026). No comparable binding framework exists in the United States.

Across these documents, the consensus commitments (human oversight, educator involvement, privacy, age-appropriateness, and transparency) are sound and match the evidence. What none of them yet require is the thing the evidence most clearly calls for: that any AI tool used with children be evaluated on learning measured *without* the tool and on relational outcomes measured at all.

Implications for Practice

The evidence does not support either enthusiasm or prohibition as a general stance. It supports a set of specific practices, which we organize by audience.

For Teachers

- **Guardrails over answers:** If students use generative AI for practice, configure or prompt it to give hints and ask questions rather than provide solutions. The one K–12 trial that tested this directly found that unguarded access reduced exam performance by 17% relative to no access, while guarded access eliminated the harm (Bastani et al., 2025). A consumer chatbot's default is to be helpful, and for a learner "helpful" often means "does the work."
- **Assess with the tool off:** Build into every unit an unassisted check of whether students can do what the AI helped them practice. Students using AI overestimate their own learning (Bastani et al., 2025); the gap between confidence and unassisted performance is itself a teaching opportunity.

- **Stay in the loop:** The largest moderator in the meta-analytic literature is teacher support during student–AI interaction (g = 1.426 with it, g = 0.077 without; Gu & Yan, 2025). AI used as an independent station produces, on the best K–12 evidence, little engagement and little learning (Oreopoulos & Low, 2026; Robinson et al., 2026). Use it as a shared object of attention or a tutor you can see over the student's shoulder.
- **Teach the metacognition the tool removes:** Make planning, monitoring, and evaluating explicit when AI is involved: predict before prompting, find the error in an AI response, explain why a hint worked. The students who benefit most from generative AI already regulate their own learning well (Zhao et al., 2025); the students who need you most do not yet.
- **Redesign rather than detect:** Cheating rates did not change with ChatGPT (Chen et al., 2026; Lee et al., 2024), and detectors are biased against multilingual writers (Liang et al., 2023). Design assignments whose cognitive work cannot be outsourced (in-class drafting, oral defense, process portfolios) and spend the time saved from policing on conferring.
- **Ask about companions:** A majority of your students have used an AI companion and a third have chosen one over a person for a serious conversation (Common Sense Media, 2025); most who use chatbots for mental health advice have told no one (McBain et al., 2026). You do not need to be a counselor to ask a student, without judgment, what they talk to the AI about and whether they also have a person. Know your referral pathway, because companion chatbots recognize a crisis far less often than you would (Brewster et al., 2025).

For School Leaders

- **Protect relational time first:** The best-evidenced use of AI in schools is to give teachers back time (Roy et al., 2024) and to make tutors better at the human work of tutoring (Wang et al., 2025). Adopt teacher-facing tools before student-facing ones, and treat the hours saved as time for relationships, not as a reason to increase caseloads.
- **Measure what you are changing:** If you adopt a student-facing AI tool, decide in advance how you will know whether students learned without it and whether anything happened to their connection to teachers and peers. Students in high-AI-use schools are more likely to report feeling less connected to their teachers and twice as likely to report companion use (Laird et

al., 2025). Those are associations, not effects, but a leader who adopts without measuring will never know which.

- **Prepare adults for the relational side:** Only about one in ten teachers has had training on responding when a student's AI use may be harming their well-being (Laird et al., 2025). Pair any rollout with professional development on recognizing dependence and on referral. Teachers' trust in AI tools rises when they understand how the tool decides and can override it (Nazaretsky et al., 2022).
- **Separate school tools from personal use:** Nearly a third of students have personal conversations with AI on school devices (Laird et al., 2025). Decide deliberately whether school-provided systems will permit open-ended personal conversation, and if they do, who is responsible for what students say in them.
- **Favor structured over open tools for struggling students:** Where the evidence shows equity-positive effects, the tool scaffolds a specific routine with an adult involved (ASSISTments, Tutor CoPilot, dialogic-reading agents). Where it shows equity-negative effects, the tool is open-ended and relies on the student's own regulation. Students in remedial settings barely use open AI tutors even when assigned to them (Oreopoulos & Low, 2026).

For Districts and Policymakers

- **Require evidence of learning without the tool:** Procurement standards should ask vendors for evidence from assessments in which the AI was unavailable. Most studies do not report this (Deng et al., 2025), and those that do find a different picture (Bastani et al., 2025). An effect size measured with the tool on is not evidence of learning.
- **Distinguish vendor evidence from independent evidence:** Vendor-commissioned reports dominate the evidence for commercial adaptive platforms and reading tutors; independent randomized trials are rare and find modest effects (Foster, 2024; Oreopoulos & Low, 2026). Weight them accordingly.
- **Prioritize guidance where use is highest:** Daily chatbot use is highest among Black and Hispanic teens (Faverio & Sidoti, 2025), and district guidance is least common in the highest-poverty schools (Kaufman et al., 2025). Direct guidance, training, and procurement support to the schools whose students are using AI most and have been told least.

- **Treat companion AI as a child-safety issue, not an edtech issue:** Companion platforms respond appropriately to adolescent crises less than a quarter of the time and provide referrals about one time in nine (Brewster et al., 2025). Age assurance, crisis-detection standards, and limits on relational design features for minors belong in child-protection policy. UNESCO's minimum age of 13 (Miao & Holmes, 2023) and Common Sense Media's recommendation against companion use under 18 (2025) are the reference points on the table.
- **Fund the research that does not exist:** There is no high-quality causal study of student AI use in U.S. K–12 classrooms, no causal study of AI's effect on teacher–student relationships or belonging, and no empirical study of AI use among foster or system-involved youth. Research funding priorities that do not name these gaps will not close them.

Limitations and Research Gaps

This review is limited by the literature it reviews. Several gaps are structural and should shape how confidently the findings above are applied.

The K–12 causal base is tiny and recent. Fewer than two dozen studies meet a credible causal standard for K–12 (Stanford SCALE Initiative, 2026), and the longest of them covers two school years. The meta-analytic literature on generative AI is 84% higher education (Deng et al., 2025). Nearly every claim about K–12 generative AI rests on extrapolation from older students, from short interventions, or from pre-generative tools.

Durability is almost never measured. Most studies measure outcomes with the tool available, or immediately after a brief intervention. The handful that measure unassisted performance (Bastani et al., 2025; Darvishi et al., 2024; Fan et al., 2025) find that gains shrink or reverse. No study we located follows K–12 students' AI-supported learning for a year or more with assessments in which the AI is absent. Higher Learning Lab's Evidence Tracker codes "durability with vs. without AI" separately for this reason; at present, the category is nearly empty.

Relational outcomes are not measured at all in causal designs. We found no randomized or quasi-experimental study of a K–12 AI tool with teacher–student relationship quality, belonging, peer relationships, or social-emotional development as an outcome. The attachment evidence is from adults, from self-selected companion users, or from cross-sectional surveys. The question of

what AI does to a child's human relationships over time is, at the moment, unanswerable from the research.

Vulnerable populations are understudied in every direction. Subgroup analyses by prior achievement exist in a few trials and point in conflicting directions. There are no studies specific to foster youth, almost none specific to students with disabilities in generative-AI contexts, and few specific to English learners beyond language-learning chatbots. Bias research has barely touched disability or socioeconomic status (Baker & Hawn, 2022).

Publication bias and retraction risk are live. Publication bias was detected in the leading meta-analysis (Deng et al., 2025); a reanalysis of another found that its effects disappeared after adjustment (Bartoš et al., 2025), and that article was retracted (Wang & Fan, 2026). Preprints with striking findings circulate widely before review (Kosmyna et al., 2025) and are sometimes contested (Stankovic et al., 2025). The incentive structure of the field favors large, fast, positive results.

Vendor and interested-party evidence is pervasive. Commercial platforms are evaluated mainly by their vendors; the strongest companion-chatbot trial was co-authored by researchers at the chatbot's maker (Fang et al., 2025); the strongest tutoring reviews come from a center dedicated to human tutoring (National Student Support Accelerator, 2026). None of this makes the findings wrong, but all of it should be visible to the reader.

The technology moves faster than the studies. Bastani et al. (2025) tested GPT-4 in the fall of 2023 and note that chatbots were then new to students; models, tutor configurations, and platform policies have all changed since. Every finding here is a finding about a particular tool at a particular moment, and the half-life of such findings is short.

What to watch. The research agenda that follows is specific. The field needs preregistered K–12 trials with unassisted outcomes at delays of months, not days; trials that measure relational and well-being outcomes alongside achievement; designs that randomize teacher involvement and guardrail configuration rather than merely access; studies that sample, or at least report on, foster youth, students with disabilities, and English learners; longitudinal cohort studies of adolescent companion use with human-relationship outcomes; independent evaluations of commercial platforms; and bias audits that include disability and socioeconomic status. Until these exist, the most honest summary is that AI *can* harm learning when unguarded, *can* help modestly when structured and supported

by adults, *does* occupy a growing share of children's emotional lives, and *has not been shown* to do anything, good or bad, to the human relationships on which children's development depends.

Conclusion

The research on artificial intelligence, learning, and attachment in K–12 education does not tell a simple story, but it does tell a coherent one. Four decades of work on intelligent tutoring systems established that computer tutors built on cognitive models of a domain can produce real learning gains, largest in mathematics, dependent on sustained implementation, and uncertain in their reach to the students who struggle most. Four years of work on generative AI has produced a large and enthusiastic literature whose headline effects come overwhelmingly from university students, over weeks, measured while the tool is still available, and whose most rigorous K–12 test found that unguarded access made high school students worse at mathematics while making them more confident. The well-designed K–12 trials since then converge on three lessons: guardrails matter, engagement is the binding constraint, and the human adult is the most powerful variable in the system. The most promising K–12 result in the generative era came from an AI that coached the tutor, not the child.

On attachment, the evidence is earlier but no less consequential. Children readily extend social and moral standing to conversational machines. Adults who use AI companions feel less lonely in the moment, and those who use them most are doing worst. A majority of American teenagers now use such systems, a third of them for relationships, and a third have chosen a machine over a person for a serious conversation. The companion platforms they use fail to recognize a crisis more than half the time. No study has yet measured what any of this does to a child's relationship with a teacher, a friend, or a caregiver, and no study has looked at the children for whom those relationships are most fragile.

For practitioners who serve vulnerable students, the implication is not to refuse the technology but to refuse the framing that treats it as a substitute for what works. The evidence supports AI that gives teachers back time, that makes tutors more skilled, that scaffolds a specific routine under a watching adult, and that is assessed by what students can do when it is turned off. The evidence does not support AI as an independent tutor for children who are already disengaged, as a confidant for children who are already alone, or as a reason to spend less adult attention on the children who need it most. The burden of proof, in a field this young and this commercially driven,

belongs with the claim that a tool helps, and the standard of proof should be learning that lasts and relationships that hold. Higher Learning Lab's Evidence Tracker exists to hold that standard in public view, study by study, as the evidence arrives.

References

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education, 21*, Article 10. https://doi.org/10.1186/s41239-024-00444-7

Ainsworth, M. D. S. (1989). Attachments beyond infancy. *American Psychologist, 44*(4), 709–716. https://doi.org/10.1037/0003-066X.44.4.709

Allen, K., Kern, M. L., Vella-Brodrick, D., Hattie, J., & Waters, L. (2018). What schools need to know about fostering school belonging: A meta-analysis. *Educational Psychology Review, 30*(1), 1–34. https://doi.org/10.1007/s10648-016-9389-8

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bartoš, F., Martinková, P., & Wagenmakers, E.-J. (2025). *Adjusting for publication bias reveals no evidence for the effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/8vs32_v1

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), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics, 3*(21), Article eaat5954. https://doi.org/10.1126/scirobotics.aat5954

Bowlby, J. (1969). *Attachment and loss: Vol. 1. Attachment*. Basic Books.

Brandtzaeg, P. B., Følstad, A., & Skjuve, M. (2025). Emerging AI individualism: How young people integrate social AI into everyday life. *Communication and Change, 1*(1), Article 11. https://doi.org/10.1007/s44382-025-00011-2

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research, 48*(3), 404–429. https://doi.org/10.1093/hcr/hqac008

Brewster, R. C. L., Zahedivash, A., Tse, G., Bourgeois, F., & Hadland, S. E. (2025). Characteristics and safety of consumer chatbots for emergent adolescent health concerns. *JAMA Network Open, 8*(10), Article e2539022. https://doi.org/10.1001/jamanetworkopen.2025.39022

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends, 66*(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y

Chen, R., Lee, V. R., Kuo, A. C., Pope, D. C., & Miles, S. (2026). Cheating in the second year of generative AI chatbots: A follow-up study on high school student cheating behaviors. *Educational Technology Research and Development, 74*(2), 639–667. https://doi.org/10.1007/s11423-026-10587-1

Common Sense Media. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions

Crawford, J., Allen, K.-A., Pani, B., & Cowling, M. (2024). When artificial intelligence substitutes humans in higher education: The cost of loneliness, student success, and retention. *Studies in Higher Education, 49*(5), 883–897. https://doi.org/10.1080/03075079.2024.2326956

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education, 210*, Article 104967. https://doi.org/10.1016/j.compedu.2023.104967

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2025). AI companions reduce loneliness. *Journal of Consumer Research, 52*(6), 1126–1148. https://doi.org/10.1093/jcr/ucaf040

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). "Hey Google is it OK if I eat you?": Initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children* (pp. 595–600). ACM. https://doi.org/10.1145/3078072.3084330

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence, 6*, Article 100230. https://doi.org/10.1016/j.caeai.2024.100230

Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature, 58*(4), 897–996. https://doi.org/10.1257/jel.20191507

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689, 12 July 2024. http://data.europa.eu/eli/reg/2024/1689/oj

European Parliament & Council of the European Union. (2026). Regulation (EU) 2026/1744 amending Regulation (EU) 2024/1689 as regards simplification and the application dates of certain provisions (Digital Omnibus on AI). *Official Journal of the European Union*. https://www.euaiact.com/digital-omnibus-ai

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). *How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study* [Preprint]. arXiv. https://arxiv.org/abs/2503.17473

Faverio, M., & Sidoti, O. (2025, December 9). *Teens, social media and AI chatbots 2025*. Pew Research Center. https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/

Festerling, J., & Siraj, I. (2020). Alexa, what are you? Exploring primary school children's ontological perceptions of digital voice assistants in open interactions. *Human Development, 64*(1), 26–43. https://doi.org/10.1159/000508499

Fitzgerald, M. (2025, July 15). More than half the states have issued AI guidance for schools. *Stateline*. https://stateline.org/2025/07/15/more-than-half-the-states-have-issued-ai-guidance-for-schools/

Foster, M. E. (2024). Evaluating the impact of supplemental computer-assisted math instruction in elementary school: A conceptual replication. *Journal of Research on Educational Effectiveness, 17*(1), 94–118. https://doi.org/10.1080/19345747.2023.2174919

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*(1), Article 42. https://doi.org/10.1007/s10648-026-10133-8

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), Article 6. https://doi.org/10.3390/soc15010006

Giovanelli, A., & Roundfield, K. D. (2025). Adolescent vulnerability to consumer chatbots—Artificial agents and genuine risk. *JAMA Network Open, 8*(10), Article e2539028. https://doi.org/10.1001/jamanetworkopen.2025.39028

Gu, J., & Yan, Z. (2025). Effects of GenAI interventions on student academic performance: A meta-analysis. *Journal of Educational Computing Research, 63*(6), 1460–1492. https://doi.org/10.1177/07356331251349620

Guan, L., Li, S., & Gu, M. M. (2024). AI in informal digital English learning: A meta-analysis of its effectiveness on proficiency, motivation, and self-regulation. *Computers and Education: Artificial Intelligence, 7*, Article 100323. https://doi.org/10.1016/j.caeai.2024.100323

Guingrich, R. E., & Graziano, M. S. A. (2024). Ascribing consciousness to artificial intelligence: Human-AI interaction and its

carry-over effects on human-human interaction. *Frontiers in Psychology, 15*, Article 1322781. https://doi.org/10.3389/fpsyg.2024.1322781

Hamre, B. K., & Pianta, R. C. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*(2), 625–638. https://doi.org/10.1111/1467-8624.00301

Hartup, W. W., & Stevens, N. (1997). Friendships and adaptation in the life course. *Psychological Bulletin, 121*(3), 355–370. https://doi.org/10.1037/0033-2909.121.3.355

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning, 38*(1), 237–257. https://doi.org/10.1111/jcal.12610

Huntington-Klein, N. (2026). *Little impact of ChatGPT availability on high school student test score performance* [Preprint]. arXiv. https://arxiv.org/abs/2605.08812

Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., Ruckert, J. H., & Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology, 48*(2), 303–314. https://doi.org/10.1037/a0027033

Kaufman, J. H., Woo, A., Eagan, J., Lee, S., & Kassan, E. B. (2025). *Uneven adoption of artificial intelligence tools among U.S. teachers and principals in the 2023–2024 school year* (Report No. RR-A134-25). RAND Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RRA100/RRA134-25/RAND_RRA134-25.pdf

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Kizilcec, R. F. (2024). To advance AI use in education, focus on understanding educators. *International Journal of Artificial Intelligence in Education, 34*(1), 12–19. https://doi.org/10.1007/s40593-023-00351-4

Korpershoek, H., Canrinus, E. T., Fokkens-Bruinsma, M., & de Boer, H. (2020). The relationships between school belonging and students' motivational, social-emotional, behavioural, and academic outcomes in secondary education: A meta-analytic review. *Research Papers in Education, 35*(6), 641–680. https://doi.org/10.1080/02671522.2019.1615116

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://arxiv.org/abs/2506.08872

Kovach, L. P. (2026). Artificial intimacy: Companion artificial intelligence and emerging risks to adolescent mental health. *Social Sciences, 15*(7), Article 491. https://doi.org/10.3390/socsci15070491

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society, 26*(10), 5923–5941. https://doi.org/10.1177/14614448221142007

Laestadius, L. I., & Campos-Castillo, C. (2026). Reminders that chatbots are not human can be risky. *Trends in Cognitive Sciences, 30*(3), 187–189. https://doi.org/10.1016/j.tics.2025.12.007

Laird, E., Dwyer, M., & Quay-de la Vallee, H. (2025). *Hand in hand: Schools' embrace of AI connected to increased risks to students*. Center for Democracy & Technology. https://cdt.org/insights/hand-in-hand-schools-embrace-of-ai-connected-to-increased-risks-to-students/

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–22). ACM. https://doi.org/10.1145/3706598.3713778

Lee, V. R., Pope, D., Miles, S., & Zárate, R. C. (2024). Cheating in the age of generative AI: A high school survey study of cheating behaviors before and after the release of ChatGPT. *Computers and Education: Artificial Intelligence, 7*, Article 100253. https://doi.org/10.1016/j.caeai.2024.100253

Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P.-M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning, 10*, Article 29. https://doi.org/10.1038/s41539-025-00320-7

Li, X., & Yow, W. Q. (2024). Younger, not older, children trust an inaccurate human informant more than an inaccurate robot informant. *Child Development, 95*(3), 988–1000. https://doi.org/10.1111/cdev.14048

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Liu, J., Mbowe, A. E., Tahri, D., & Aziku, M. (2026). Meta-analysis on the influence of AI agents on K-12 student cognitive performance. *Computers in Human Behavior Reports, 21*, Article 100973. https://doi.org/10.1016/j.chbr.2026.100973

Lo, C. K., Hew, K. F., & Jong, M. S.-Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education, 219*, Article 105100. https://doi.org/10.1016/j.compedu.2024.105100

Lovato, S. B., & Piper, A. M. (2019). Young children and voice search: What we know from human-computer interaction research. *Frontiers in Psychology, 10*, Article 8. https://doi.org/10.3389/fpsyg.2019.00008

Ma, N., & Zhong, Z. (2025). A meta-analysis of the impact of generative artificial intelligence on learning outcomes. *Journal of Computer Assisted Learning, 41*(5), Article e70117. https://doi.org/10.1111/jcal.70117

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. https://doi.org/10.1037/a0037123

Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj*

Mental Health Research, 3, Article 4. https://doi.org/10.1038/s44184-023-00047-6

Marino, M. T., Vasquez, E., Dieker, L., Basham, J., & Blackorby, J. (2023). The future of artificial intelligence in special education technology. *Journal of Special Education Technology, 38*(3), 404–416. https://doi.org/10.1177/01626434231165977

Marzano, D. (2025). Generative artificial intelligence (GAI) in teaching and learning processes at the K-12 level: A systematic review. *Technology, Knowledge and Learning, 31*(2), 789–829. https://doi.org/10.1007/s10758-025-09853-7

McBain, R. K., Cantor, J. H., Breslau, J., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Pataranutaporn, P., Stein, B. D., Mehrotra, A., & Yu, H. (2026). AI chatbot use and disclosure for mental health among US adolescents and young adults. *JAMA Pediatrics, 180*(8), 884–890. https://doi.org/10.1001/jamapediatrics.2026.2015

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Mosher, M. A., Kohnke, S., Holman, K., Sulaimon, T., Dieker, L., & Frey, B. B. (2026). Intervention fidelity for enhanced outcomes: A systematic review of artificial intelligence in special education. *Journal of Special Education Technology*. Advance online publication. https://doi.org/10.1177/01626434261464947

Mostow, J., Nelson-Taylor, J., & Beck, J. E. (2013). Computer-guided oral reading versus independent practice: Comparison of sustained silent reading to an automated reading tutor that listens. *Journal of Educational Computing Research, 49*(2), 249–276. https://doi.org/10.2190/EC.49.2.g

Murphy, R., Roschelle, J., Feng, M., & Mason, C. A. (2020). Investigating efficacy, moderators and mediators for an online mathematics homework intervention. *Journal of Research on Educational Effectiveness, 13*(2), 235–270. https://doi.org/10.1080/19345747.2019.1710885

Namvarpour, M., Chang, T., & Razi, A. (2026). *A responsive present, a shared past, a social other: Teens' overreliance on companion AI chatbots* [Preprint]. arXiv. https://arxiv.org/abs/2609.14843

National Student Support Accelerator. (2026). *AI tutoring is not a monolith: What we actually know*. Stanford University, SCALE

Initiative. https://nssa.stanford.edu/sites/default/files/AI%20Tutoring%20is%20Not%20a%20Monolith%20What%20We%20Actually%20Know.pdf

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology, 53*(4), 914–931. https://doi.org/10.1111/bjet.13232

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal, 61*(1), 74–107. https://doi.org/10.3102/00028312231208687

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

Oreopoulos, P., & Low, N. (2026). *One click away: AI tutoring with Khanmigo in a two-year school experiment* (NBER Working Paper No. 35620). National Bureau of Economic Research. https://doi.org/10.3386/w35620

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at scale. *Educational Evaluation and Policy Analysis, 36*(2), 127–144. https://doi.org/10.3102/0162373713507480

Regan, P. M., & Jesse, J. (2019). Ethical challenges of edtech, big data and personalized learning: Twenty-first century student sorting and tracking. *Ethics and Information Technology, 21*(3), 167–179. https://doi.org/10.1007/s10676-018-9492-2

Robinson, C. D., Gormley, D., Ribeiro, A. T., & Loeb, S. (2026). *Access is not enough: Human support improves engagement with AI tutoring* (EdWorkingPaper No. 26-1451). Annenberg Institute at Brown University. https://edworkingpapers.com/ai26-1451

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*(4), 493–529. https://doi.org/10.3102/0034654311421793

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open, 2*(4). https://doi.org/10.1177/2332858416673968

Roy, P., Poet, H., Staunton, R., Aston, K., & Thomas, D. (2024). *ChatGPT in lesson preparation: A Teacher Choices Trial*. National Foundation for Educational Research / Education Endowment Foundation. https://www.nfer.ac.uk/publications/chatgpt-in-lesson-preparation-a-teacher-choices-trial/

Sidoti, O., Park, E., & Gottfried, J. (2025, January 15). *About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in 2023*. Pew Research Center. https://www.pewresearch.org/short-reads/2025/01/15/about-a-quarter-of-us-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion – A study of human-chatbot relationships. *International Journal of Human-Computer Studies, 149*, Article 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies, 168*, Article 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

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, 160*, Article 108386. https://doi.org/10.1016/j.chb.2024.108386

Stanford SCALE Initiative. (2026, March 11). *The evidence base on AI in K-12: A 2026 review*. Stanford Graduate School of Education. https://scale.stanford.edu/research-in-action/understanding-evidence-base-ai-k12-education

Stankovic, M., Hirche, E., Kollatzsch, S., & Doetsch, J. N. (2025). *Comment on: Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing tasks* [Preprint]. arXiv. https://arxiv.org/abs/2601.00856

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105*(4), 970–987. https://doi.org/10.1037/a0032447

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research, 22*(3), Article e16235. https://doi.org/10.2196/16235

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–24). ACM. https://doi.org/10.1145/3613904.3642902

U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*. U.S. Department of Education.

U.S. Department of Education. (2025, July 22). *U.S. Department of Education issues guidance on artificial intelligence use in schools, proposes additional supplemental priority* [Press release]. https://www.ed.gov/about/news/press-release/us-department-of-education-issues-guidance-artificial-intelligence-use-schools-proposes-additional-supplemental-priority

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wang, J., & Fan, W. (2025). RETRACTED ARTICLE: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*, Article 621. https://doi.org/10.1057/s41599-025-04787-y

Wang, J., & Fan, W. (2026). Retraction Note: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 13*, Article 528. https://doi.org/10.1057/s41599-026-07310-z

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://arxiv.org/abs/2410.03017

Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). ChatGPT's impact on student learning outcomes: A meta-analysis of 35 experimental studies. *Humanities and Social Sciences*

Communications, 13, Article 684. https://doi.org/10.1057/s41599-026-07019-z

Xie, T., & Pentina, I. (2022). Attachment theory as a framework to understand relationships with social chatbots: A case study of Replika. In *Proceedings of the 55th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2022.258

Xu, Y., Aubele, J., Vigil, V., Bustamante, A. S., Kim, Y.-S., & Warschauer, M. (2022). Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Development, 93*(2), e149–e167. https://doi.org/10.1111/cdev.13708

Xu, Y., Vigil, V., Bustamante, A. S., & Warschauer, M. (2022). "Elinor's talking to me!": Integrating conversational AI into children's narrative science programming. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. https://doi.org/10.1145/3491102.3502050

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yang, F., & Oshio, A. (2025). Using attachment theory to conceptualize and measure the experiences in human-AI relationships. *Current Psychology, 44*(11), 10658–10669. https://doi.org/10.1007/s12144-025-07917-6

Zhao, Y., Yue, Y., Sun, Z., Jiang, Q., & Li, G. (2025). Does generative artificial intelligence improve students' higher-order thinking? A meta-analysis based on 29 experiments and quasi-experiments. *Journal of Intelligence, 13*(12), Article 160. https://doi.org/10.3390/jintelligence13120160