Learning outcomes with GenAI in the classroom

A review of empirical evidence

AETHER AI ETHICS AND EFFECTS IN ENGINEERING AND RESEARCH

Executive summary

This report presents a review of recent empirical evidence of generative AI (GenAI) impact on learning outcomes in formal education. Its purpose is to provide educators with an overview of top concerns for ensuring students' learning gains when using LLM-based learning tools and concludes with research-derived guidance for deciding when and how to use these tools in the classroom. The report unfolds as follows:

Section 1 distinguishes between the needs of education and industry, where the benefits of LLMs were first explored, primarily for productivity gains. Educators' priorities are different. Pedagogical concerns include consideration of inequities in education, developing students' critical thinking skills, and the potential for GenAI to inhibit social development. These concerns extend beyond technologists' focus on mitigating technical harms such as toxic content, bias, or accuracy in system outputs.

Section 2 presents several key variables that affect learning with GenAI: (1) AI literacy—understanding the capabilities and limitations of an AI system—is a critical new variable for student success when using GenAI. **(2)** Educational equity is a variable where GenAI renders mixed experiences for marginalized groups. Studies show how GenAI can be an effective resource for students with disabilities. In other contexts, it entrenches existing patterns in academic performance of the weakest students and can exacerbate inequities for economically marginalized students. **(3)** GenAI can impact psychological and social conditions long recognized to facilitate learning: self-efficacy, individual pace, and human connection. On self-efficacy, studies show that students can be overconfident about their skill mastery when using GenAI and need help calibrating their mental model of learning gains. For self-paced learning, GenAI introduces both efficiencies and pitfalls depending on learning domain and context, including whether AI tools are general purpose chatbots or scaffolded tutors. Studies also highlight GenAI impact on human connection, the foundation for developing higher-order skills of critical thinking and creativity. GenAI's on-demand availability but lack of social presence can present opportunities and disadvantages, from providing a nonjudgmental environment for exploring topics to reducing collaboration with peers in group projects. Yet, studies show that human tutors remain students' preferred source for trusted information.

Section 3 examines how GenAI usage aligns with learning objectives in Bloom's taxonomy. Basic cognitive skills—Bloom's *remembering* and *understanding*—are fundamental to success across academic domains. Studies show that there can be an overdependence and lack of engagement that result in impaired memory

formation when using LLM chatbots. Development of higher-order thinking—*analysis, reasoning,* and *creativity*—can be compromised if GenAI is used in ways that bypass the necessary struggle that is integral to acquiring skills. Studies illustrate how use of general-purpose GenAI tools such as ChatGPT, without scaffolding or other pedagogical guardrails, can be detrimental to critical thinking. GenAI can also impact creativity. Students using GenAI for creative problem-solving can benefit from fast prototype iteration and greater project completeness or detail but can also tend toward idea fixation and less originality and complexity in their work.

Section 4 highlights how GenAI learning tools need greater pedagogical complexity. Up to now, state-of-the-art tools have been ChatGPT or similar, with prompt engineering for the model to assume an instructor role or restrain its outputs. However, modified general-purpose chatbots cannot address the broad range of pedagogical considerations involved in learning success. New types of experimental AI tutors with embedded proven pedagogical strategies—for example, capable of detecting and effectively responding to a range of student cognitive states—show promise. Consulting educators in the design is key for success of systems like these that are on the horizon.

A concluding synthesis of the empirical evidence offers four guidelines for integrating GenAI in learning environments: (1) Ensure student readiness—avoid introducing GenAI too early, before students master domain basics. **(2)** Teach AI literacy—build an awareness of GenAI capabilities and limitations so students can assess system outputs and learn domain-specific techniques for optimal results. **(3)** Use GenAI as a supplement to traditional learning methods—GenAI explanations and examples are capabilities that students value, but teacher guidance with these explanations remains necessary. **(4)** Promote design interventions that foster student engagement—limiting copy-paste functionality, supporting students' metacognitive calibration to reduce overestimation of their learning progress, nudging learners towards critical thinking, and evaluating GenAI tools for proven engagement strategies.

Authors

Kathleen Walker
Responsible AI
Research Associate

Mihaela Vorvoreanu, PhD
Principal Applied Scientist
Co-chair, Aether Psi

Acknowledgments: This report is an effort of the Microsoft Aether Psychological Influences of AI (Psi) working group. We thank Psi co-chairs Jina Suh, PhD, and Forough Poursabzi, PhD, for their review and extend a special thanks to Lev Tankelevitch, PhD, and Advait Sarkar, PhD, for their contribution.

Introduction

Large language models' (LLMs') probabilistic capacity to generate outputs in human-like utterances is often framed as a societally disruptive technology, reframing work and relevant skills for many white-collar professions. Today's workforce is already expected to be able to use AI technologies as they progress from classroom to careers, according to Microsoft's 2025 Work Trend Index (Microsoft WorkLab, 2025). School administrators and educators are under pressure to integrate AI tools in the classroom.

With ChatGPT's public release in November 2022, we are still historically early in GenAI usage in the classroom. Understandably, we do not yet have many large-scale studies nor any long-term data of its effect on learning outcomes. Plus, there are few K-12 studies. Meanwhile, student and faculty sentiment varies greatly, with many caveats, including concerns regarding GenAI's impact on critical thinking and human deskilling. The known flaws in GenAI system outputs—biases, confabulations, and tendency toward homogeneity—along with multiple alarming news media reports (Ropek, 2025) further compound difficult decision-making around how to effectively use the technology in classrooms.

The goal of this report is to help guide educators in navigating and making informed decisions around GenAI's education opportunities, such as scalable learning through on-demand, personalized AI tutors. For busy educators, this report provides an overview of empirical evidence of GenAI impact on student learning outcomes.

As recent experiments illustrate, learning outcomes with GenAI tools are nuanced and context dependent. Yet, consistent themes emerge, which, when distilled, can help inform educators and education technology (edtech) designers for best practice.

This report is structured in four main sections:

1. **How the education paradigm differs from industry,** where automation has traditionally been a boost.
2. **How conditions for learning are influenced by GenAI,** including a new need for AI literacy, aspects of educational equity, and psychosocial factors such as self-efficacy, individual pace, and human connection.
3. **How learning objectives of Bloom's taxonomy are impacted by GenAI**—from basic cognitive skills of remembering and understanding to higher-order thinking skills of analysis and creativity.
4. **How GenAI tools should more closely reflect the complexities of pedagogy,** as illustrated in recent experimental AI tutors.

It concludes by synthesizing the summarized evidence into high-level guidance for educators wishing to integrate GenAI into learning effectively: ensuring student readiness, teaching AI literacy, using GenAI as a supplement, promoting interventions that encourage engagement, and following proven pedagogy in GenAI tools.

To start, it is helpful to clarify distinctions between the goals of education and industry, where GenAI must meet separate aims.

1 Education paradigm differs from industry

"Preparation for the workforce" is frequently cited as the purpose of current education. However, despite formal education being vital to healthy industry, the immediate goals of these two entities differ. In industry, productivity is the goal, a measurable standard that new automation technologies can boost. For example, GenAI's textual automation has been shown to increase workers' productivity in different contexts (e.g., Dell'Acqua et al., 2023). Formal education's goal is different from productivity. It is "to equip learners with knowledge or skills that are both durable and flexible" (Soderstrom & Bjork, 2015). Therefore, speed, quantity, and ease of accomplishing tasks are more important in the workplace than in the classroom, where learning entails a level of necessary struggle.

1.1 Productivity does not equate to learning gains

GenAI's productivity improvements in industry might not be desirable in education settings. In a survey of knowledge workers, from music directors to medical staff, participants using GenAI for a variety of tasks reported increased efficiency but reduced the perceived enaction of critical thinking, particularly when users had greater confidence in AI for that task (Lee et al., 2025a). This outcome is distinctly at odds with formal education's goal of developing thinking skills. For example, in coding classes, students may cover more topics in less time with GenAI, but "at the cost of a shallower understanding" (Lehmann et al., 2025). High school math students have been shown to complete more practice exercises when using GenAI; but when later tested, they scored 17% lower than the control group who never had access (Bastani et al., 2024). Students report time pressure as a factor in their use of GenAI. There's evidence suggesting that the heavier the student's workload or the more they procrastinate, the more they rely on ChatGPT and experience impaired recall (Abbas et al., 2024). While time savings can be a primary motivation for students using GenAI tools (K. Wang et al., 2024), high productivity in educational settings should not be confused with skill mastery.

1.2 Educators have complex pedagogical concerns

Understanding how to deploy GenAI in the classroom in ways that support learning objectives clearly differs from how industry harnesses the technology for automation benefits. Yet educators are often not consulted in the design or purchase of AI systems despite their having to comply with mandated usage. For example, researchers identified that only eight out of 29 LLM-based systems proposed from 2020 to 2023 for expressly supporting education—from answer-grading apps to virtual tutors—had included either teachers or students in design or development (Garcia-Mendez et al., 2024).

In interviews with educators and leading edtech providers, Harvey et al. (2025) found that edtech providers' focus is quite different from teachers'. Technologists concentrate on the immediacy of mitigating technical harms such as toxic content, bias, or accuracy in their system outputs. Educators, while sharing these concerns, point to more complex issues at stake—the

potential for GenAI to exacerbate inequities within education, impede critical thinking skills, and inhibit social development.

2 Learning conditions: Key variables in learning with GenAI

The potential for GenAI to facilitate education at scale is a promising aspiration. However, there is a lack of consensus across the research literature as to the effectiveness of GenAI in learning performance. Furthermore, what is meant by *learning performance* or *outcomes* across studies is widely varied, as are the research methods and experiment conditions. Many early studies focused on student motivation and engagement. But while affective states are integral to learning, they are not measures of acquired skill. The result is, as noted in an analysis of 48 studies, "a significant gap in validating GenAI's effectiveness in measuring and enhancing learning outcomes" (X. Zhang et al., 2024). Other analyses of studies range from finding a small but positive effect of GenAI on learning performance (Zhu et al., 2025) to reporting a large overall positive effect (e.g., Deng et al., 2024; Wang & Fan, 2025). Dissenting voices caution against rushing GenAI implementation in the classroom, with research findings either of no statistically meaningful impact (Thoeni & Fryer, 2025) or detrimental effects (Kosmyna et al., 2025).

As always, context is everything. For example, Deng et al. (2024), in an analysis of 69 experimental studies between 2022 and 2024 assessing ChatGPT's impact on student learning, note a predominant focus on language education (31.88% of the studies' subject areas). Although they find a large overall positive effect on learning performance, they recommend caution when interpreting studies' results, given, for example, that measures of higher-order thinking were self-reports.

Wang and Fan (2025) further clarify the importance of context. While researchers saw a large positive effect of ChatGPT on learning performance in 44 of 51 experimental studies published November 2022 to February 2025, they identified several moderating variables. These include the type of course (e.g., STEM, language learning), learning approach (e.g., personalized to an individual pace; problem-based learning for real-world solutions), and how ChatGPT is used (e.g., scaffolded or intelligent tutoring, instant feedback). The authors are clear that there are many other potential variables to be examined, including sociocultural and economic factors.

Learning success with AI has many interdependencies. In this section, we examine empirical evidence about key variables that impact learning with GenAI: AI literacy, equity, self-efficacy, individual learning pace, and the requirement for human connection—each crucial in the development of thinking skills.

2.1 AI literacy: GenAI introduces a critical new variable

Throughout studies, the need for AI literacy—understanding the capabilities and limitations of GenAI—is voiced by students and apparent in experimental data (e.g., X. Zhang et al., 2024;

Atcheson et al., 2025; Kim et al., 2024; Kreijkes et al., 2025). Whether students are equipped with appropriate strategies for using GenAI can make a difference in learning outcomes. Research shows that novices have difficulty with effective prompting (e.g., Urban et al., 2024) and that students do not always understand the advantage of iterative interaction. Students with the greatest interactive engagement and iteration—reading, re-prompting, and editing the output, as well as using GenAI for brainstorming—perform better than those who limit their use to copy-and-pasting responses or using GenAI as an information source (Nguyen et al., 2024).

Graduate students who received in-depth AI literacy instruction—not only information about the system, but also hands-on demonstrations of techniques for GenAI co-authoring—were 64.5% faster and had a grade improvement from B+ to A when completing professional memo-writing tasks that required citing and evaluating sources for different audiences (Usdan et al., 2024). While this experiment's results may appear to highlight productivity and goals distinctive from academic learning, it illuminates benefits of in-depth AI literacy instruction. Such instruction can be particularly helpful in learning domains like computer science, where GenAI is deeply integrated.

Education about how to use GenAI should reflect the experiences of their student populations and not merely "deliver knowledge" (Cao et al., 2025). In a study of hands-on activities with students in an underserved K-12 district with limited STEM opportunity, Cao et al. (2025) urge using the power of cultural capital for teaching AI literacy. When students built their own AI models, created self-portraits with GenAI, and found questions ChatGPT could not answer correctly, they could contextualize AI capabilities as well as its biases and errors within their own firsthand discoveries.

2.2 Educational equity: GenAI renders mixed experiences

In one context, GenAI can be an effective self-advocacy resource for students with disabilities. University students with disabilities report that using GenAI has helped improve accessibility in learning materials and increased their independence for finding academic and community resources (Atcheson et al., 2025). In a different context, research shows that GenAI can further entrench existing patterns in academic performance. For example, in high school math, it is the weakest students' learning outcomes that are most harmed when GenAI tools are deployed without pedagogical guardrails (Bastani et al., 2024). In a study of English language learning in Nigeria, those who gained the most benefit from GenAI access were higher academic achievers to begin with (De Simone et al., 2025).

2.2.1 GenAI can exacerbate inequities for economically marginalized students

Educators express concern about GenAI exacerbating persistent inequities, including for economically marginalized students (Harvey et al., 2025). There is basis for this concern. For example, in a study of 947 U.S. college students using ChatGPT, researchers found higher family socioeconomic status (SES) was associated with greater student AI literacy—their knowledge of the technology's capabilities and limitations. Students with higher family SES used ChatGPT more for both academic support, such as asking for explanations or brainstorming, and for non-

recommended uses, like getting summaries without reading an assignment. Students with lower SES, such as financially burdened first-generation students, demonstrated lower digital and AI literacy. These students were more inclined to over-trust outputs and experience a diminished sense of engagement (C. Zhang et al., 2024).

There is some evidence GenAI can be effective in low economic resource settings. De Simone et al. (2025), in a six-week pilot program with Nigerian students, demonstrated potential for cost-effective LLM English-language tutoring, a skill with economic impact for the students' future careers. But there was still an SES differentiator: students with lower household SES made the least learning gains.

GenAI does not radically disrupt the traditional pattern of SES in academic performance. Yu et al. (2024) examined whether there was an equitizing effect of LLM-based tools on the readability and sophistication of student writing since public access to LLMs. Analyzing more than 1.1 million writing submissions on U.S. college course forums since 2021, they observed a decrease in the gap between linguistically advantaged and disadvantaged students' work, particularly for non-native English speakers. However, this equitizing effect was more concentrated among students with higher SES.

Similar concerns for equity are raised when surveying educators, developers, and literature on GenAI impact on competencies in university computer science classrooms, a domain where the technology is now integral to the profession (Prather et al., 2025). Barriers to access or accomplishing goals with available AI tools can include low literacy, not being a native English speaker, or socioeconomics. The question of who can afford the subscription versus the free version of GenAI tools can be a factor in widening the gap between who is better prepared for career entry.

Equity in education is a complex social issue, one that cannot be addressed singlehandedly by GenAI learning tools. As illustrated, access may improve specific groups' learning conditions while for others it may not budge the needle toward equity.

2.3 Psycho-social conditions in learning

Empirical evidence shows that psychological and social conditions long recognized to facilitate learning, such as self-efficacy, individual pace, and human connection, remain central to discussions of GenAI impact in education. We see that GenAI can mislead students about their true learning gains, offer both efficiencies and pitfalls for self-paced learning, and risk reducing teacher-student interaction that is vital to motivation and critical thinking.

2.3.1 Self-efficacy: Overconfidence with GenAI

Self-efficacy, or one's confidence in their ability to accomplish a task, is a complex factor in learning, with many dependencies and mixed effects when using GenAI. Mixed effects include some students gaining and others losing confidence when comparing their abilities with GenAI outputs (Simkute et al., 2025).

Self-efficacy is integral to student success. Researchers find computer science students with a history of higher grades and self-efficacy scores accelerate with GenAI tools as they are less prone to being distracted by unhelpful GenAI outputs (Prather et al., 2024). Students with high academic self-efficacy tend to use GenAI tools more for explanations and as a supplemental rather than primary resource (C. Zhang et al., 2024).

But studies show GenAI can distort students' perception, facilitating an overestimation of their actual performance. For example, in Urban et al.'s creative problem-solving experiment where students were tasked with improving a toy, higher self-efficacy was associated with increased overestimation of performance in both the control and experiment group. However, self-efficacy was demonstrably greater among those using ChatGPT. Further, the more useful participants claimed ChatGPT to be, greater the overestimation of their work's quality and originality.

Students need help calibrating their mental model of learning gains. A faulty mental model of how we learn is a common phenomenon, and one that leads students to mismanage their learning (Bjork et al., 2013). In a large-scale study of high school math students using GenAI, it is the weakest students who were most harmed by this false sense of confidence (Bastani et al., 2024). GenAI can exacerbate this phenomenon particularly for novices in a learning domain (Singh et al., 2025). Students who are in early stages of learning a new skill may misunderstand initial instructions, go through steps too quickly, and misinterpret their GenAI use as enhancing their skills (Prather et al., 2024). Lehmann et al. (2025) found in experiments with students learning Python that LLM usage was most detrimental for students who started with the least domain knowledge. Furthermore, after working with an LLM, their assessment of their skills exceeded the reality of their graded final assignment.

For better learning outcomes with GenAI, students need help calibrating their mental model of learning gains. To this end, Lee al. (2025b) experimented with a real-time AI-powered tool to help students better predict their actual knowledge acquisition. The tool intervened at 15-minute intervals throughout an assigned study topic, prompting participants to compare their self-evaluation with an AI-predicted score of their learning. Students could then identify and focus on areas where they were weak. While there was significant increase in mean learning gains for those using the AI support tool (16.3% over the control group's 7.4%), there was also a 4.1% improvement in the metacognitive calibration of students who were initially overconfident.

As studies clarify, GenAI can foster a false sense of confidence in students' actual learning. Interventions that help students form an appropriate mental model of their learning gains are core to benefits GenAI may offer.

2.3.2 Self-paced learning: Efficiencies and pitfalls with GenAI

There are many variables in self-paced learning with GenAI. These include whether AI tools are general purpose chatbots or scaffolded tutors and what the learning domain may be. When evaluating claims of GenAI efficiencies in learning scenarios, we are reminded of essential differences between efficiency and proficiency. For example, research shows that what participants perceive as efficiencies when writing an essay—ChatGPT expediting research and

feedback—do not necessarily hold over time, as memory and recall are impaired (Kosmyna et al., 2025).

An experiment by Kestin et al. (2024) demonstrates there can be measurable efficiencies for true learning with appropriately designed AI tutors. Mindful of pedagogical best practices, they developed an online GenAI tutor to study its effect in a Harvard undergraduate physics course. During the first week, one group received an AI-tutored lesson at home and the other did the same lesson by attending an in-person active-learning session with an instructor. In the second week, the groups swapped learning venues. Findings showed that students working with the custom-designed AI tutor at home were able to pace themselves, learning twice as much in less time on task and reporting an increased sense of engagement and motivation. Researchers emphasize that replicating this outcome on a broader scale would require careful engineering of AI systems that reflect pedagogical best practices, such as scaffolding.

Zamfirescu-Pereira et al. (2025) examined the time-savings efficiency of a GPT-4 AI assistant for more than 2,000 students and their teaching staff across two semesters in a Berkeley Computer Science 1 course. They designed a bot that assessed the students' knowledge based on their submitted code, then provided hints but not solutions. With the bot, students in the 50th to 80th percentile completed homework assignments up to 50% faster than the previous year's cohort. Researchers did not observe any appreciable decrease in student grades when compared with the same course in prior years without an AI homework assistant. However, they did observe a 75% drop in homework-related questions posted to the course's online forum, where students traditionally seek help from human teaching assistants. They raise an important question of AI tutors introducing tradeoffs with human connection that is important in learning support.

2.3.3 Human connection: GenAI impact on a prerequisite to higher-order skills

Learning is a sociocultural process, where human connection is the foundation for developing higher-order thinking skills (Vygotsky & Cole, 1978). GenAI can challenge this connection when, for example, asking a TA for help entails a wait time but a bot responds immediately. GenAI's constant availability has the potential to supplant relationships with people as well as other resources. Undergraduate students cite GenAI as their first go-to for information before consulting other search tools, friends, or teachers (Simkute et al., 2025). While GenAI is valued as a safe place for students to explore topics (e.g., Atcheson et al., 2025), human tutors are preferred by students for trusted information. The teacher-student connection remains a relationship crucial to fostering engagement and critical thinking. A survey of 230 university students across academic disciplines shows a preference for human tutors because of their perceptive guidance and emotional connection, despite students crediting ChatGPT with enabling self-paced learning, giving instant feedback, and providing a nonjudgmental space (Fakour & Imani, 2025).

Social presence is a catalyst for engagement and learning and throughout research there are traces where students feel shortchanged by GenAI. For example, non-native English speakers using a ChatGPT-embedded writing system express frustration with the system's inability to

detect writers' intent or cultural context, sensibilities that are expected of a human teacher (Kim et al., 2025). This lack of human presence is considered a possible detractor in students' attitudes toward learning with ChatGPT across a variety of studies that otherwise report overall large positive effects of GenAI (Wang & Fan, 2025).

In collaborative projects with peers—an important dimension of human connection in learning— there's a tendency for students to pull back and work solo when they have GenAI access (Albadarin et al., 2024). Both the absence of and withdrawal from human connection have implications for higher-order skills of critical thinking and creativity.

3 Learning objectives: Examining how GenAI usage aligns with Bloom's taxonomy

There's a long history of initial skepticism and fears about technologies that change human processes, such as the printing press or calculator, to which people and society satisfactorily adapted. However, GenAI may be different than earlier unsettling technologies. Singh et al. (2025) point across literature to GenAI's unique capacity to disrupt reflexivity, citing its harmful impacts of echo chambers, cognitive offloading, reduced critical thinking, and impressive outputs that "create an illusion of comprehensive understanding" for the learner. The framework of progressively ordered and interdependent thinking skills outlined in Bloom's taxonomy of learning objectives (Krathwohl, 2002) helps parse how GenAI usage can affect student outcomes. As seen in the following empirical evidence, GenAI can compromise students' thinking, from basic recall and comprehension to higher-order skills of analysis and creativity.

3.1 Basic cognitive skills—remembering and understanding

Retention and comprehension are fundamental to success across academic domains and are the earliest of learning objectives in Bloom's taxonomy—*remembering* and *understanding*. There is evidence that GenAI can curtail these learning basics.

An investigation of LLM chatbot impact on reading comprehension and memory retention clarifies that traditional learning approaches are far from obsolete. Kreijkes et al. (2025) measured the effects of traditional note-taking versus working with an LLM chatbot in active reading assignments with 344 high school students. Students were divided into two groups and randomly assigned two passages, each less than 400 words, on the history of apartheid and the Cuban missile crisis. For each group, one passage was an LLM-only condition, where students did not take notes but instead explored the topic with the chatbot, getting explanations and other support, like generating a quiz. Group 1, for their second reading, made notes with no chatbot access. In contrast, Group 2 had the LLM chatbot available when making notes (LLM+Notes), including copy-paste functionality—which more than 25% used heavily. Three days later students were tested for their retention, comprehension, and free recall. Traditional note-taking outperformed LLM+Notes across all three measures, with the LLM-only condition

scoring the lowest. Notably, for free recall, not only was traditional note-taking significantly more effective, but LLM+Notes showed no substantial benefit over the LLM-only condition.

In a similar vein, Kosmyna et al. (2025) found LLM use can impair memory formation and recall as they examined how LLMs' impact during essay writing. Recruiting university students, they divided participants into three groups, each assigned to exclusively use the assistance of an LLM (ChatGPT), a search engine, or no tools—a "Brain-only" group. During several sessions, the groups wrote argumentative essays in 20 minutes, selecting topics from SAT prompts. When interviewed immediately following the essay exercise, LLM users' recall was impaired—they were unable to quote from the essays they had just written. In contrast, the Brain-only group's memory recall was remarkably greater, along with their reporting deeper satisfaction and ownership of their work. The LLM users underperformed the Brain-only group significantly in linguistic quality and scoring by both human teachers and AI.

Factors of overdependence and lack of engagement surface in LLM users' behavior. This has import not only for memory retention but also for higher-order thinking skills.

3.2 Higher-order thinking skills

Learning requires cognitive grappling. Development of higher-order thinking—analysis, reasoning, and creativity—can be compromised if GenAI is used in ways that bypass the "desirable difficulties" (Bjork & Bjork, 2011) that are integral to acquiring skills. For example, research shows that greater learning gains in math are made by those who first try solving on their own before LLM assistance (Kumar et al., 2023). Educators and their students who use GenAI as a tutor, writing assistant, or brainstorming partner voice concerns about overdependence and losing their own skills, especially critical thinking (Simkute et al., 2025).

3.2.1 Impact of cognitive offloading

There are the stubborn challenges in improving student learning outcomes, including factors of motivation, different learning paces, and high student-teacher ratios. Throughout the early claims of GenAI's potential to overcome these hurdles, there was much talk about GenAI's benefit for cognitive offloading, keeping students focused and not overwhelmed by extraneous information. However, as the experiment measuring recall of essay writers using LLM assistance shows, reducing cognitive load with GenAI in learning and creative processes carries the risk of *cognitive debt* (Kosmyna et al., 2025). There's concern that offloading mental effort in the short term may have long-term costs for critical thinking and creativity. Stadler et al. (2024) found lower cognitive load associated with reduced quality of argumentation and reasoning in essays that were researched and written by students using ChatGPT compared with students limited to Google Search.

But perhaps one of the most concerning findings of the risks of excessive cognitive offloading is exemplified in Kreijkes et al.'s note-taking experiment, where major historical events were the assigned topics: out of the students' more than 4000 prompts, only 6 questioned the correctness of ChatGPT's outputs.

3.2.2 Impact of general-purpose GenAI on critical thinking

Using general-purpose GenAI tools such as ChatGPT for learning without scaffolding or other explicit pedagogical guardrails can be detrimental to critical thinking. A review of learning trends in computer science presents the conundrum of AI tools reducing students' analytical thinking while learning to code at a time when knowing how to critically evaluate GenAI outputs is in itself a newly required competency (Prather et al., 2025). Researchers advise against using general-purpose tools such as ChatGPT, as they "widen the gap" in critical thinking skills in code problem-solving for novice programmers (Prather et al., 2024).

A systematic review of 14 studies focused on university students' reliance on AI chatbots found that while AI conversational agents may help with speeding up tasks such as research and information retrieval, these tools frequently reduced students' critical thinking and reasoning skills (Zhai et al., 2024). Students in these studies expressed concern not only about GenAI's potential for confabulation and other technical flaws, but also an awareness of their own potential deskilling and loss of creativity as tradeoffs for speed and convenience.

Xue et al. (2024) raise the question whether excessive reliance on AI reduces explorations that are essential to creative problem-solving. In an experiment with an introductory programming class, half the students had access to online resources such as course slides, Google, and Stack Overflow, while the other half also used ChatGPT. While researchers found that ChatGPT usage did not result in any statistically significant impact on learning outcomes, they observed that students using ChatGPT largely did not access any other educational resources, such as web pages or course slides.

3.2.3 GenAI impact on creativity and creative problem-solving

The concern about GenAI narrowing students' use of other resources and reducing their creativity and problem-solving skills is legitimate. The paradigm of "use it or lose it" for intellectual skills is paramount at a time when the world needs creative solutions for problems assaulting the environment and civil society (Sternberg, 2024). We see greater variety and divergence across human responses in standardized creativity tests when compared with the significant similarity of responses generated by a broad set of LLMs (Wenger & Kenett, 2025). GenAI is limited to reconstituting training data. This has implications for reducing divergent thinking when people use LLMs for brainstorming and creative problem-solving. This potential impact on divergent thinking and creativity is further compounded when we see students choosing to work individually with GenAI rather than collaborate with peers (Alabadarin et al., 2024).

Sandhaus et al. (2024) found that brainstorming, along with reflecting on solutions, can be compromised in creative problem-solving. Students in an applied human-computer interaction course documented their permitted and self-initiated use of GenAI tools (e.g., GPT-5, DALL-E) in a project designing interactive devices. Findings showed that brainstorming during use-case development and storyboarding suffered, as students reduced their exploration and fixated on ideas. There was also evidence that participants could find GenAI to be a consciously

constraining experience: when interested in a task that required "deeper involvement," students often chose to work without GenAI. For instance, a participant described the benefits of physically drawing when exploring solutions, finding their mind engaged in a more efficient and enjoyable way than prompting a chatbot.

In contrast, Urban et al. (2024) suggest that ChatGPT offers the advantage of exposing learners to a broad range of potential prototypes and support for iterating, helping students develop better solutions for creative problem-solving. In their experiment, university students were assigned a toy improvement task with the goal of outcompeting a competitor's sales. Researchers measured the quality (how well the project's defined goals were met), degree of detail, and the originality of student's work. They found that ChatGPT strongly influenced the quality of solutions and had a moderate impact on detail and originality.

Project completeness vs. originality and complexity

Interestingly, we see reports of increased project completeness or detail with GenAI in creative learning scenarios as opposed to a large effect on originality or sophistication (e.g., Urban et al., 2024). Study participants are frank—they get "co-pilot to fill in with [the] long parts" of labs and documentation (Sandhaus et al., 2024).

Completeness over creativity is characteristic of GenAI support. When examining creativity and argumentation in academic writing, Niloy et al. (2024) found ChatGPT improved essay elaboration and presentation, but not originality. In their experiment with 600 university students writing a timed essay, the control group had access to Google Search while the experimental group also had ChatGPT. Their essays were both machine- and human-evaluated for a total creativity score. Evaluation took into account the originality (vs. similarity with AI-generated content); content presentation (including readability and visual content); accuracy (appropriate and reliable details); and elaboration (refining and expanding ideas, narrative flow). The experimental group's essays with ChatGPT indicated a significant drop in overall quality from their pre-test mean scores. As for their improved elaboration and presentation, researchers suggest that ChatGPT's ability to give easy-to-understand explanations may have been a boost.

4 Increased pedagogical complexity is needed in GenAI learning tools

Modifying general-purpose chatbots as AI tutors can be too simplistic, incapable of addressing the range of pedagogical considerations for learning success. Up to now state-of-the-art tools have generally been ChatGPT or similar, with prompt engineering for the model to assume role of instructor and restraining generation with other basic commands, such as what topic and teaching strategy to use, restricting responses from giving solutions, or requesting quizzes for testing knowledge (e.g., Mollick & Mollick, 2023). Verifying outputs is an expected part of this

terrain. Reviewing 29 educational GenAI systems, researchers found that less than a third provided enough transparency about their code and data to verify claims of accuracy (Garcia-Mendez et al., 2024). Both edtech providers and educators are cognizant that "if LLMs cannot be made to adhere to evidence-based teaching approaches," proving their contribution to learning will be difficult (Harvey et al., 2025).

There's a growing recommendation to avoid problematic general-purpose tools like ChatGPT in favor of scaffolded GenAI tools for improving learning outcomes (e.g., Prather et al., 2024; Kestin et al., 2024). Some take a more radical stance: when a custom-designed retrieval-augmented generative (RAG) AI tutor using Socratic method was surprisingly shown to have no statistically significant impact on interest, self-efficacy, engagement, or test performance across 450 university students, researchers recommended halting funds until more is understood for effectively harnessing GenAI (Thoeni & Fryer, 2025).

New types of experimental tutors better reflect complexities of pedagogy. Consulting educators in the design is key. Zha et al. (2025), with design input from several educators, engineered an experimental GenAI tutor that mentored a dozen students, ages 12 to 14, through a creative problem-solving process as they completed tasks like creating smart home devices. The system detected and responded to a range of student behavior and cognitive states—e.g., *silent, shallow thinking, inadequate information integration skills,* or *fatigue*—applying guidance from 20 proven pedagogical strategies. Researchers found this AI coaching system significantly improved student engagement, creativity, and task performance when compared with a baseline system representative of current state-of-the-art tutors using chatbots with minimal prompt engineering, where students tended to prompt for direct solutions and copy-paste.

Scaling expertise remains a constant stated goal for AI tutors. Some research is exploring ways this might be achieved without compromising human connection that is fundamental to students' intellectual development. Promising areas include an experimental AI tutor helping human tutors improve their skills with pedagogical strategies in real time during live online math tutoring (R. Wang et al., 2025). An AI system that simulates disengaged students can help train teachers to apply strategies for online learning success (Pan et al., 2025).

Meanwhile as research unfolds for better AI tools with embedded pedagogy, educators must reconcile usage mandates with the current technology at their disposal. To help support educators' decision-making, this report concludes with a set of basic guidelines derived from the empirical evidence of learning outcomes with GenAI.

Conclusion: Basics for benefiting from GenAI learning support

Conditions across these recent investigations of GenAI impact on learning outcomes vary. Experiments include choosing to use unrestricted LLMs or placing simple guardrails (e.g.,

prompting chatbot for instructor role and not providing direct solutions), using AI tutors with designed UI interventions, and a concentration of participants in specific domains (i.e., computer science) or education levels (i.e., primarily tertiary, few K-12). The many nuances prevent a pure comparison of apples to apples. Nevertheless, we can distill high-level guidance for approaching design and deployment of GenAI in the classroom. These may be described as basics for benefiting from GenAI as learning support: ensuring student readiness, teaching AI literacy, using GenAI as a supplement, promoting interventions that encourage engagement, and following proven pedagogy in GenAI tools.

Learning with GenAI requires student readiness

Student readiness is a key consideration for improving learning outcomes with GenAI tools, reducing opportunity for GenAI to foster a false sense of confidence. Readiness also includes students' ability to critically assess GenAI outputs, a key competency in student AI literacy.

Don't introduce GenAI too early. Premature introduction can impair comprehension and basic recall, harm the novice, and foster an illusion of learning. Research suggests withholding GenAI during early phases of learning for memory formation (Kosmyna et al., 2025) and recommends maintaining traditional learning strategies that support comprehension and retention (Kreijkes et al., 2025). The value of *desirable difficulties*—the cognitive grappling essential in learning—should not be undermined by AI tools.

Domain novices first need mastery of the basics in their new field, or run the risk of shallow engagement and understanding, plus a false sense of confidence in their skills (e.g., K. Wang et al., 2024; Zamfirescu-Periera et al., 2025; Sandhaus et al., 2024; Lehmann et al., 2025). For example, without a solid grasp of domain basics, coding novices tend to misunderstand initial instructions and rush through steps, misunderstanding their GenAI usage as improving their skills (Prather et al., 2024).

Teaching AI literacy is foundational to gaining GenAI benefits

AI literacy is essential to education's goal of developing higher-order skills. There are many components to AI literacy, as well as competencies that students must develop.

Building awareness of GenAI capabilities and limitations is a fundamental for equipping students to critically assess outputs. Students must understand GenAI potential for confabulation and biases, its summarization strengths and weaknesses, and the need to verify outputs, cross-referencing with other sources. Throughout studies, the need for this awareness is directly voiced by students and apparent in experimental data (e.g., X. Zhang et al., 2024; Atcheson et al., 2025; Kim et al., 2024; Kreijkes et al., 2025).

Learning hands-on, domain-specific techniques for optimal results is a core competency for AI literacy. When students are equipped with strategies for prompting and iterative interaction with LLMs, their performance is stronger (Nguyen et al., 2024). How-to demonstrations of domain-specific techniques can help optimize student-LLM interactions (Usdan et al., 2024).

Know your audience when designing for AI literacy. To be effective, learning materials should reflect the experiences of their student populations and not merely "deliver knowledge" (Cao et al., 2025). Codesigning AI education materials with the students they are meant to reach can be highly effective, as Cao et al. demonstrated in K-12 students' hands-on discovery of capabilities, limitations, errors, and biases when creating their own AI models and other activities.

Use as a supplement: Learning from explanations and examples

GenAI should be a supplement to, not a replacement for, traditional learning methods. General-purpose tools like ChatGPT can interfere with learning processes, from recall to critical thinking. However, GenAI, despite flaws, can provide clear and helpful explanations of complex concepts, a capability students value. Surveying STEM students, K. Wang et al. (2024) report that students find GenAI to be "most helpful for explaining relevant domain knowledge underlying a problem and least helpful for verifying solution correctness." Coding students who use LLMs for explanations, as a complement rather than a solution provider, deepen their understanding (Lehmann et al., 2025). As Kreijkes et al. (2025) clarify in their note-taking experiment, there is a recommended sequence when recall and retention are objectives: students should first read and make notes on their own before exploring topics with LLMs.

Teacher guidance with GenAI explanations remains necessary. For example, Shao et al. (2025) in an experiment with high school students found that generating simple analogies—like *dominoes* for explaining *chain reactions* or imagining the body as a city and the immune system as its police force in explaining *immune response*—can effectively improve student understanding of domain terminology. However, the catch is that teachers needed to adjust the analogies for correctness and completeness to be optimally effective, as well as guide students to avoid overdependence.

Promote interventions that foster engagement

How students use GenAI, whether for explanations or directly getting solutions, makes all the difference in learning outcomes. Small changes in the way GenAI tools are designed can have a big impact.

Limit copy-paste functionality. Increasing the transaction cost encourages memory formation and gives opportunity for critical evaluation (e.g., Lehmann et al., 2025).

Support metacognitive calibration. Students overestimating their learning progress is a recurring challenge with GenAI. Providing regular interventions, such as intermittent quizzes, can help students have a clear picture of their domain knowledge and where they need to focus (Lee et al., 2025b).

Nudge for critical thinking. Learners should be prompted to pause and reflect, consider other points of view, and gauge how well they understand a GenAI output. Metacognitive prompts at critical points in GenAI interaction have potential to help students develop an awareness of their own thinking (Singh et al., 2025).

Evaluate GenAI tools for proven pedagogy. Be aware that modifying general-purpose chatbots like ChatGPT with prompts cannot accommodate complex factors in learning, such as student motivation and engagement. Some researchers recommend avoiding this type of AI tutor for early domain learning (e.g., Prather et al., 2024), while advocating for careful engineering of AI systems that reflect pedagogical best practices (Kestin et al., 2024). Consulting educators in the design of such systems is key, with few system providers having done this historically (Garcia-Mendez et al., 2024). Carefully engineered systems that are designed in collaboration educators can incorporate a wide array of pedagogical strategies to better reflect the realities of student learning. For example, a single AI tutor that is designed to match states of student engagement and levels of understanding to pedagogical strategies for responses can yield significantly better outcomes than GenAI systems that are only prompt-engineered (Zha et al., 2025).

Although many well-designed AI tutors may be on the horizon, integrating GenAI into the classroom remains a balancing act of overlapping considerations. Educators must consider GenAI's impact on thinking skills and its efficiencies and pitfalls for self-paced learning alongside risks for domain novices or students who struggle. As GenAI adoption in the classroom moves forward, two conditions are critical for success: AI literacy and human connection. As students indicate (Fakour & Imani, 2025), there is no trusted equal to the teacher-student relationship for evaluating information and developing higher-order thinking—the ultimate skills for the after-school world.

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(1), 10.

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. Discov Educ3 60

Atcheson, A., Khan, O., Siemann, B., Jain, A., & Karahalios, K. (2025, April). "I'd Never Actually Realized How Big An Impact It Had Until Now": Perspectives of University Students with Disabilities on Generative Artificial Intelligence. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-22).

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, O., & Mariman, R. (2024). Generative ai can harm learning. Available at SSRN, 4895486.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68), 56-64.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, *64*(1), 417-444.

Cao, H. J., Choi, K., Park, C., & Lee, H. R. (2025, April). AI Literacy for Underserved Students: Leveraging Cultural Capital from Underserved Communities for AI Education Research. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality**.** *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies**.** *Computers & Education*, 105224.

De Simone, M. E., Tiberti, F. H., Rodriguez, M. R. B., Manolio, F. A., Mosuro, W., & Dikoru, E. J. (2025). *From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in nigeria* (No. 11125). The World Bank.

Fakour, H., & Imani, M. (2025). Socratic wisdom in the age of AI: a comparative study of ChatGPT and human tutors in enhancing critical thinking skills. In *Frontiers in Education* (Vol. 10, p. 1528603). Frontiers Media SA

García-Méndez, S., de Arriba-Pérez, F., & Somoza-López, M. D. C. (2024). A review on the use of large language models as virtual tutors. *Science & Education*, 1-16.

Harvey, E., Koenecke, A., & Kizilcec, R. F. (2025, April). "Don't Forget the Teachers": Towards an Educator-Centered Understanding of Harms from Large Language Models in Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2024). AI tutoring outperforms active learning.

Kim, J., Yu, S., Detrick, R., & Li, N. (2025). Exploring students' perspectives on Generative AI-assisted academic writing. Education and Information Technologies, 30(1), 1265-1300.

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. *arXiv preprint arXiv:2506.08872*, *4*.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212-218.

Kreijkes, P., Kewenig, V., Kuvalja, M., Lee, M., Vitello, S., Hofman, J. M., Sellen, A., Rintel, S., Goldstein, D. G., Rothschild, D. M., Tankelvitch, L., & Oates, T. (2025). Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. Available at SSRN.

Kumar, H., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Math education with large language models: peril or promise?. *Available at SSRN 4641653*.

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*. Association for Computing Machinery, New York, NY, USA, Article 1121, 1–22.

Lee, H., Stinar, F., Zong, R., Valdiviejas, H., Wang, D., & Bosch, N. (2025, April). Learning Behaviors Mediate the Effect of AI-powered Support for Metacognitive Calibration on Learning Outcomes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Lehmann, M., Cornelius, P. B., & Sting, F. J. (Mar 2025). AI Meets the Classroom: When Do LLMs Harm Learning? arXiv preprint arXiv:2409.09047.

Leppänen, L., Aunimo, L., Hellas, A., Nurminen, J. K., & Mannila, L. (2025). How Large Language Models Are Changing MOOC Essay Answers: A Comparison of Pre-and Post-LLM Responses.

Microsoft WorkLab. (2025). *Work Trend Index Annual Report*. https://www.microsoft.com/en-us/worklab/work-trend-index/2025-the-year-the-frontier-firm-is-born#section-intelligence-on-tap

Mollick, E. R., & Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *The Wharton School Research Paper*

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. Studies in Higher Education, 49(5), 847-864.

Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2024). Is Chatgpt a menace for creative writing ability? An experiment. Journal of computer assisted learning, 40(2), 919-930

Pan, S., Schmucker, R., Garcia Bulle Bueno, B., Llanes, S. A., Albo Alarcón, F., Zhu, H., Teo, A. & Xia, M. (2025, April). Tutorup: What if your students were simulated? Training tutors to address engagement challenges in online learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Prather, J., Leinonen, J., Kiesler, N., Gorson Benario, J., Lau, S., MacNeil, S., Norouzi, N., Opel, S., Pettit, V., Porter, L., Reeves, B. N., Savelka, J., Smith IV, D. H., Strickroth, S., & Zingaro, D. (2025). Beyond the Hype: A Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools. *2024 Working Group Reports on Innovation and Technology in Computer Science Education*, 300-338.

Prather, J., Reeves, B. N., Leinonen, J., MacNeil, S., Randrianasolo, A. S., Becker, B. A., Kimmel, B., Wright, J., & Briggs, B. (2024, August). The widening gap: The benefits and harms of generative ai for novice programmers. In Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1 (pp. 469-486)

Ropek, L. (2025, May 16). *It's Breathtaking How Fast AI Is Screwing Up the Education System*. Gizmodo. https://gizmodo.com/its-breathtaking-how-fast-ai-is-screwing-up-the-education-system-2000603100

Sandhaus, H., Gu, Q., Parreira, M. T., & Ju, W. (2024). Student Reflections on Self-Initiated GenAI Use in HCI Education. arXiv preprint arXiv:2410.14048.

Shao, Z., Yuan, S., Gao, L., He, Y., Yang, D., & Chen, S. (2025, April). Unlocking Scientific Concepts: How Effective Are LLM-Generated Analogies for Student Understanding and Classroom Practice?. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Simkute, A., Kewenig, V., Sellen, A., Rintel, S., & Tankelevitch, L. (2025). The New Calculator? Practices, Norms, and Implications of Generative AI in Higher Education. arXiv preprint arXiv:2501.08864.

Singh, A., Taneja, K., Guan, Z., & Ghosh, A. (2025). Protecting Human Cognition in the Age of AI. arXiv preprint arXiv:2502.12447.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176-199.

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*, 108386.

Sternberg, R. J. (2024). Do not worry that generative AI may compromise human creativity or intelligence in the future: It already has. Journal of Intelligence, 12(7), 69

Thoeni, A., & Fryer, L. K. (2025). AI Tutors in Higher Education: Comparing Expectations to Evidence.

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. Computers & Education, 215, 105031.

Usdan, J., Connell Pensky, A., & Chang, H. (2024). Generative AI's Impact on Graduate Student Writing Productivity and Quality. *Available at SSRN*.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Wang, J., & Fan, W. (2025). 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*(1), 1-21.

Wang, K. D., Wu, Z., Tufts II, L. N., Wieman, C., Salehi, S., & Haber, N. (2024). Scaffold or Crutch? Examining College Students' Use and Views of Generative AI Tools for STEM Education. *arXiv preprint arXiv:2412.02653*.

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024). Tutor CoPilot: A human-AI approach for scaling real-time expertise. arXiv preprint arXiv:2410.03017.

Wenger, E., & Kenett, Y. (2025). We're Different, We're the Same: Creative Homogeneity Across LLMs. *arXiv preprint arXiv:2501.19361*.

Xue, Y., Chen, H., Bai, G. R., Tairas, R., & Huang, Y. (2024, April). Does ChatGPT help with introductory programming? An experiment of students using ChatGPT in CS1. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training* (pp. 331-341 CS1).

Yu, R., Xu, Z., CH-Wang, S., & Arum, R. (2024). Whose ChatGPT? Unveiling Real-World Educational Inequalities Introduced by Large Language Models. arXiv preprint arXiv:2410.22282.

Zamfirescu-Pereira, J. D., Qi, L., Hartmann, B., DeNero, J., & Norouzi, N. (2025, February). 61A Bot Report: AI Assistants in CS1 Save Students Homework Time and Reduce Demands on Staff.(Now What?). In Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1 (pp. 1309-1315).

Zha, S., Liu, Y., Zheng, C., Xu, J., Yu, F., Gong, J., & Xu, Y. (2025, April). Mentigo: An Intelligent Agent for Mentoring Students in the Creative Problem Solving Process. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-22).

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, *11*(1), 28.

Zhang, C., Rice, R. E., & Wang, L. H. (2024). College students' literacy, ChatGPT activities, educational outcomes, and trust from a digital divide perspective. New Media & Society, 14614448241301741.

Zhang, X., Zhang, P., Shen, Y., Liu, M., Wang, Q., Gašević, D., & Fan, Y. (2024). A Systematic Literature Review of Empirical Research on Applying Generative Artificial Intelligence in Education. Frontiers of Digital Education, 1(3), 223-245.

Zhu, Y., Liu, Q., & Zhao, L. (2025). Exploring the impact of generative artificial intelligence on students' learning outcomes: a meta-analysis. *Education and Information Technologies*, 1-29.