can LLMs be used to improve student learning?

LLMs can improve student learning when used for scaffolding comprehension, generating practice questions, or providing feedback on student work, but they may impair learning when they substitute for cognitively demanding practice, making their effectiveness dependent on task type, structured implementation, and human educator oversight.

Abstract

This systematic review of 12 studies examining LLM applications in education finds that LLMs can improve student learning, but effects are highly task-dependent and contingent on implementation approach. Positive outcomes consistently emerged when LLMs were used for retrieval practice question generation (16-percentage-point improvement in quiz accuracy) , feedback on written work (d = 0.19 for revision performance) , and explanation generation for comprehension support . LLM-assisted instruction also improved motivational outcomes including task motivation (d = 0.36) and positive emotions (d = 0.34) . However, when LLMs substituted for cognitively demanding practice, negative effects emerged: significant negative correlations were found between LLM use for code generation and debugging and final grades (ρ = -0.305 and ρ = -0.360, respectively) , and students using LLMs for research tasks demonstrated lower-quality reasoning than those using traditional search engines .

The evidence indicates that LLMs are most effective when they scaffold understanding or provide feedback rather than replace effortful practice, when students engage actively with LLM outputs rather than passively consuming them , and when human educators provide structured guidance and verify LLM-generated content . Students without prior LLM experience showed greater performance gains than experienced users , and concerns about hallucinations and potential reinforcement of misconceptions underscore the continued importance of human oversight. In summary, LLMs represent a promising but not universally beneficial educational tool, with effectiveness determined primarily by whether implementations support rather than supplant the cognitive effort required for learning.

Paper search

We performed a semantic search using the query ”can LLMs be used to improve student learning?” across over 138 million academic papers from the Elicit search engine, which includes all of Semantic Scholar and OpenAlex.

We retrieved the 500 papers most relevant to the query.

Screening

We screened in sources based on their abstracts that met these criteria:

- **Large Language Model Intervention**: Does the study explicitly use Large Language Models (e.g., GPT models, BERT, T5, Claude, Gemini, or other transformer-based language models) as an educational tool or intervention?
- **Student Population**: Does the study involve students/learners at any educational level (K-12, undergraduate, graduate, professional training, or continuing education) in formal or informal educational settings?
- **Learning Outcomes Measurement**: Does the study include quantitative or qualitative measures of student learning (such as test scores, skill assessments, comprehension measures, knowledge retention, or academic performance)?
- **Empirical Data or Systematic Review**: Does the study present original empirical data (experimental, quasi-experimental, observational, or mixed-methods) OR is it a systematic review/meta-analysis that synthesizes evidence on LLMs and student learning?

- **Publication Quality**: Is the study published in a peer-reviewed journal or available as a high-quality preprint with clear methodology?
- **LLM Technology Focus**: Does the study focus on Large Language Models rather than solely examining traditional AI, machine learning algorithms, or other educational technologies that are not LLMs?
- **Empirical Evidence**: Does the study contain empirical data rather than being purely theoretical, conceptual, editorial, or opinion-based without data?
- **Educational Outcomes Connection**: Does the study connect LLM performance or capabilities to educational outcomes or student learning measures, rather than only evaluating technical performance (accuracy, speed, computational efficiency)?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **LLM Application**:

  Extract detailed information about how the LLM was used, including:

- Specific LLM application (e.g., feedback generation, question creation, research assistant, reflection guide)
- Which LLM was used (e.g., GPT-3.5, GPT-4, ChatGPT)
- Level of guidance/scaffolding provided to students
- Integration method (e.g., embedded in learning platform, standalone tool, instructor-mediated)
- Frequency and duration of LLM use

- **Educational Context**:

  Document the learning environment and participants:

- Educational level (e.g., secondary, undergraduate, graduate)
- Subject area/discipline
- Sample size and key demographic characteristics
- Course setting (e.g., classroom, online, hybrid)
- Task or assignment type where LLM was used

- **Study Design**:

  Extract study methodology details:

- Study design type (e.g., RCT, quasi-experimental, observational)
- Comparison condition(s)
- Duration of intervention/observation period
- Randomization method if applicable
- Key methodological strengths or limitations

- **Learning Outcomes**:

  Identify all learning-related outcomes measured, including:

- Academic performance measures (test scores, grades, assignment quality)
- Cognitive outcomes (knowledge retention, critical thinking, problem-solving)

- Motivational/affective outcomes (engagement, confidence, emotions)
- Learning process measures (effort, time spent, cognitive load)
- How each outcome was measured (assessment method, timing)

- **Key Findings**:

  Extract primary results regarding LLM effects on learning:

- Direction of effects (positive, negative, null) for each outcome
- Effect sizes with confidence intervals when reported
- Statistical significance levels
- Magnitude of differences between groups
- Any differential effects based on task type or student characteristics

- **Mechanisms/Explanations**:

  Document explanations for why LLMs helped or hindered learning:

- Proposed causal mechanisms from authors
- Qualitative findings about student experiences
- Mediating factors identified
- Trade-offs observed (e.g., efficiency vs. depth)
- Author interpretations of unexpected findings

- **Moderating Factors**:

  Extract factors that influenced LLM effectiveness:

- Task characteristics that enhanced/reduced benefits
- Student characteristics associated with better/worse outcomes
- Implementation factors that mattered
- Contextual variables affecting results
- Boundary conditions or limitations identified

Characteristics of Included Studies

The review encompasses 12 studies examining LLM applications in educational contexts, spanning multiple disciplines, educational levels, and LLM use cases.

Study	Full text retrieved?	Study Design	Educational Level	Subject Area	Sample Size	LLM Application
Yuan An et al., 2025	Yes	Quasi-experimental	Undergraduate	Data Science	~60 students	Question generation for retrieval practice
Gregor Jošt et al., 2024	Yes	Quasi-experimental	Undergraduate	Software Development (React)	32 students	Informal use for code generation, debugging, explanations

Study	Full text retrieved?	Study Design	Educational Level	Subject Area	Sample Size	LLM Application
Ahmad A. Bany Abdelnabi et al., 2025	Yes	Mixed-methods observational	Undergraduate (M3 medical)	Medical Education	100 students	Feedback on H&P tasks
Harshit Kumar et al., 2025	No	RCT	Not specified	Mathematics	1,200 participants	Explanation generation
Danielle R. Thomas et al., 2025	Yes	RCT	Undergraduate	Tutor Training	885 tutors	Feedback generation
Wenhan Lyu et al., 2024	Yes	Quasi-experimental	Undergraduate	Computer Science	50 students	Programming assistant (CodeTutor)
Harsh Kumar et al., 2023	Yes	Mixed-methods RCT	Undergraduate	Computer Science	145 (classroom) + 356 (Prolific)	Chatbot tutor
Matthias Stadler et al., 2024	No	RCT	University	Socio-scientific inquiry	91 students	Research assistant
Harsh Kumar et al., 2024	Yes	Randomized field experiments	Undergraduate	Computer Science	145 (Study 1) + 112 (Study 2)	Reflection guide
Jennifer Meyer et al., 2023	No	RCT	Upper secondary	English as Foreign Language	459 students	Feedback generation
Juho Leinonen et al., 2023	Yes	Quasi-experimental	Undergraduate	Computer Science	~1,000 students	Code explanation generation
Ze-Min Liu et al., 2024	No	RCT	Elementary	EFL Writing	65 students	SRL strategy instruction support

The studies employed diverse methodological approaches, with six RCTs , five quasi-experimental designs , and one mixed-methods observational study . Sample sizes ranged considerably from 32 students to 1,200 participants . Computer science and programming education represented the most common domain (five studies), followed by language learning (two studies), with the remainder spanning medical education, mathematics, tutoring, and scientific inquiry. Most LLM applications utilized GPT-3 or GPT-3.5 variants , with some studies employing GPT-4 or Google AI Studio .

Effects of LLMs on Learning Outcomes

Academic Performance

Study	Outcome Measure	Direction	Effect Size/Magnitude	Statistical Significance
Yuan An et al., 2025	Quiz accuracy	Positive	16 percentage points (89% vs 73%)	$p < 0.0001$
Gregor Jošt et al., 2024	Final grades (code generation)	Negative	$\rho = -0.305$	$p = 0.045$
Gregor Jošt et al., 2024	Final grades (debugging)	Negative	$\rho = -0.360$	$p = 0.021$
Gregor Jošt et al., 2024	Final grades (explanations)	Null	$\rho = -0.201$	$p = 0.135$
Harshit Kumar et al., 2025	Test performance	Positive	Not reported	Not reported
Danielle R. Thomas et al., 2025	Posttest scores (2 of 7 lessons)	Positive	$d = 0.28$, $d = 0.33$	$p < 0.05$
Wenhan Lyu et al., 2024	Final scores	Positive	+12.50 (experimental) vs -3.17 (control)	$p = 0.009$
Harsh Kumar et al., 2023	Task performance	Marginally positive	Not reported	Not significant
Harsh Kumar et al., 2024	Exam scores	Positive trend	Not reported	$p = 0.1008$ (non-significant)
Jennifer Meyer et al., 2023	Revision performance	Positive	$d = 0.19$	Not reported
Ze-Min Liu et al., 2024	Writing performance	Positive	Not reported	Significant

The majority of studies examining direct academic performance found positive effects, though the magnitude and consistency varied considerably. The strongest effects emerged when LLMs were used for structured retrieval practice, yielding a 16-percentage-point improvement in quiz accuracy . LLM-generated feedback for writing tasks produced small but positive effects on revision quality ($d = 0.19$) , and semester-long use of an LLM programming assistant resulted in significant score improvements . However, two of seven lessons in the tutor training study showed significant effects while five did not , indicating substantial variability across content areas.

Cognitive and Reasoning Outcomes

Study	Outcome	Direction	Key Finding
Gregor Jošt et al., 2024	Problem-solving skills	Negative	Overreliance on LLMs for code generation hindered independent problem-solving
Ahmad A. Bany Abdelnabi et al., 2025	Critical thinking	Positive	Enhancement in critical thinking and case analysis skills

Study	Outcome	Direction	Key Finding
Matthias Stadler et al., 2024	Quality of reasoning	Negative	Lower-quality reasoning and argumentation despite reduced cognitive load
Wenhan Lyu et al., 2024	Critical thinking	Negative	CodeTutor less effective in enhancing critical thinking skills
Juho Leinonen et al., 2023	Code comprehension	Positive	LLM explanations rated more accurate (56% vs 44%) and easier to understand (60% vs 40%)

A notable divergence emerged regarding cognitive outcomes. While LLMs effectively supported comprehension and explanation tasks , multiple studies found that reliance on LLMs for cognitively demanding tasks such as code generation, debugging, and scientific reasoning was associated with reduced quality of independent thinking . Students using LLMs as research assistants demonstrated lower-quality reasoning in their final recommendations compared to those using traditional search engines .

Motivational and Affective Outcomes

Study	Outcome	Direction	Effect Size
Harshit Kumar et al., 2025	Perceived learning	Positive	Increased perceived learning, decreased perceived difficulty
Harsh Kumar et al., 2024	Self-confidence	Positive	Significant increase ($p = 0.046$)
Jennifer Meyer et al., 2023	Task motivation	Positive	$d = 0.36$
Jennifer Meyer et al., 2023	Positive emotions	Positive	$d = 0.34$
Ze-Min Liu et al., 2024	Writing motivation	Positive	Significant improvement maintained at follow-up
Harsh Kumar et al., 2023	Trust in LLMs	Negative over time	Significant decline after task completion

Affective outcomes were consistently positive across studies, with LLM-generated feedback increasing task motivation ($d = 0.36$) and positive emotions ($d = 0.34$) , and LLM-guided reflection increasing self-confidence . However, one study noted that trust in LLM responses declined significantly after students gained practical experience with the tools , and students developed a growing preference for human teaching assistants over time .

Learning Process Measures

Study	Process Measure	Finding
Matthias Stadler et al., 2024	Cognitive load	Significantly lower with LLM use
Danielle R. Thomas et al., 2025	Completion time	No significant increase with LLM feedback
Harsh Kumar et al., 2023	Query quality	Structured guidance reduced random queries and copy-pasting

LLMs consistently reduced cognitive burden during information gathering tasks without significantly increasing time on task . The quality of student interactions with LLMs improved when structured pedagogical guidance was provided .

Synthesis: Reconciling Heterogeneous Findings

The divergent findings across studies can be explained through systematic analysis of task type, implementation approach, and student engagement patterns.

Task-Dependent Effects

The clearest pattern across studies is that LLM effectiveness depends critically on the nature of the learning task. LLMs consistently produced positive outcomes when used for:

- **Retrieval practice and knowledge testing** : A 16-percentage-point improvement in quiz accuracy occurred when LLMs generated multiple-choice questions for retrieval practice
- **Explanation and comprehension support** : LLM-generated explanations were rated significantly more accurate and easier to understand than student-generated ones , and participants who received LLM explanations for math problems showed improved learning
- **Feedback generation** : Writing feedback from LLMs improved revision performance and tutor training posttest scores in specific lessons

Conversely, LLMs produced negative effects when used for:

- **Code generation and debugging** : Significant negative correlations emerged between LLM use for these tasks and final grades (ρ = -0.305 for code generation, ρ = -0.360 for debugging)
- **Independent reasoning tasks** : Students using ChatGPT for research showed lower-quality reasoning than those using traditional search engines

This pattern suggests that LLMs enhance learning when they scaffold understanding or provide feedback on student work, but may impair learning when they substitute for cognitively demanding practice that builds foundational skills .

The Role of Engagement and Student Agency

A critical moderating factor across studies is whether students actively engaged with LLM outputs. In the tutor training study, learners with higher propensity to engage with LLM feedback scored significantly higher at posttest , and the effectiveness of feedback was contingent on learners actively seeking and using it . Similarly, the benefits of LLM explanations were largest when participants attempted problems on their own before consulting LLM explanations , though benefits persisted even when explanations were provided first .

Students without prior experience with LLM-powered tools demonstrated significantly greater performance gains than experienced users (increase of 18.877 vs. general experimental group improvement of 12.50) , suggesting a novelty effect or that experienced users may develop counterproductive reliance patterns.

Human Guidance as a Critical Implementation Factor

Studies that embedded LLMs within structured pedagogical frameworks consistently reported more positive outcomes than those examining unstructured LLM use. When teachers provided structured guidance strategies—including examples, metacognitive questioning, and solve-then-refine approaches—students showed reduced random queries and copy-pasting of assignment questions . The CALLA-LLM model, which integrated LLMs into an existing instructional framework with explicit human educator involvement, produced significant improvements in writing performance, self-regulated learning strategy use, and motivation that were maintained at one-month follow-up .

The ”Humans in the Loop” approach emphasizes the essential role of human educators in AI-assisted instruction , and multiple authors noted that instructor verification and curation of LLM outputs remained necessary . Without such structure, the efficiency gains from LLMs may come at the cost of learning depth .

Differential Effects by Domain and Task Complexity

The content complexity and alignment between LLM capabilities and learning objectives influenced outcomes. In the tutor training study, lessons on motivation showed stronger effects from LLM feedback than lessons on equity, possibly due to content complexity or feedback alignment differences . The effectiveness of LLM feedback also varied by the type of feedback provided—elaborated explanatory feedback yielded larger learning effects than simple correctness feedback .

In programming education, LLMs proved effective for syntax comprehension but ineffective for developing critical thinking skills . Students with majors in data science, mathematics, and biology showed higher final scores than computer science majors when using the LLM programming assistant , suggesting that disciplinary background interacts with LLM effectiveness in ways that warrant further investigation.

Quality and Accuracy Concerns

Several studies identified LLM limitations that constrained their educational value. Hallucinations and inconsistencies in LLM responses were observed in medical education applications , and there were concerns about LLMs potentially reinforcing misconceptions during reflection activities . The quality of LLM-generated questions varied, requiring instructor verification before deployment . The accuracy of LLM responses was significantly correlated with the quality of user prompts , indicating that student prompt-crafting skills mediate LLM effectiveness.

Conclusions by Context

Based on this synthesis, LLMs can improve student learning under specific conditions:

Most effective contexts:

- Generating practice questions and retrieval activities for knowledge consolidation
- Providing explanatory feedback on written work or open-ended responses
- Supporting comprehension through worked examples and explanations
- Facilitating structured self-reflection when embedded in pedagogical frameworks

Potentially counterproductive contexts:

- Substituting for cognitively demanding practice in skill-building domains
- Unstructured use for code generation or problem-solving without human guidance
- As primary research tools for developing argumentation and reasoning skills

Critical implementation requirements:

- Human educator involvement in curating, verifying, and contextualizing LLM outputs
- Structured pedagogical guidance that scaffolds productive LLM interactions
- Design that encourages student effort before LLM consultation
- Attention to task type alignment with LLM capabilities

References

Ahmad A. Bany Abdelnabi, Bulent Soykan, Danish Bhatti, and G. Rabadi. "Usefulness of Large Language Models (LLMs) for Student Feedback on H&P During Clerkship: Artificial Intelligence for Personalized Learning." *ACM Transactions on Computing for Healthcare*, 2025.

Danielle R. Thomas, Conrad Borchers, Shambhavi Bhushan, Erin Gatz, Shivang Gupta, and Ken Koedinger. "LLM-Generated Feedback Supports Learning If Learners Choose to Use It." *EC-TE*, 2025.

Gregor Jošt, Viktor Taneski, and Sašo Karakatič. "The Impact of Large Language Models on Programming Education and Student Learning Outcomes." *Applied Sciences*, 2024.

Harsh Kumar, Ilya Musabirov, Mohi Reza, Jiakai Shi, Xinyuan Wang, J. Williams, Anastasia Kuzminykh, and Michael Liut. "Guiding Students in Using LLMs in Supported Learning Environments: Effects on Interaction Dynamics, Learner Performance, Confidence, and Trust." *Proc. ACM Hum. Comput. Interact.*, 2023.

Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, et al. "Supporting Self-Reflection at Scale with Large Language Models: Insights from Randomized Field Experiments in Classrooms." *ACM Conference on Learning @ Scale*, 2024.

Harshit Kumar, David Rothschild, Daniel G. Goldstein, and Jake M Hofman. "Math Education With Large Language Models: Peril or Promise?" *International Conference on Artificial Intelligence in Education*, 2025.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. "Using LLMs to Bring Evidence-Based Feedback into the Classroom: AI-Generated Feedback Increases Secondary Students' Text Revision, Motivation, and Positive Emotions." *Computers and Education: Artificial Intelligence*, 2023.

Juho Leinonen, Paul Denny, S. Macneil, Sami Sarsa, Seth Bernstein, Joanne Kim, Andrew Tran, and Arto Hellas. "Comparing Code Explanations Created by Students and Large Language Models." *Annual Conference on Innovation and Technology in Computer Science Education*, 2023.

Matthias Stadler, M. Bannert, and Michael Sailer. "Cognitive Ease at a Cost: LLMs Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry." *Computers in Human Behavior*, 2024.

Wenhan Lyu, Yimeng Wang, Tingting Chung, Yifan Sun, and Yixuan Zhang. "Evaluating the Effectiveness of LLMs in Introductory Computer Science Education: A Semester-Long Field Study." *ACM Conference on Learning @ Scale*, 2024.

Yuan An, John Liu, Niyam Acharya, and Ruhma Hashmi. "Enhancing Student Learning with LLM-Generated Retrieval Practice Questions: An Empirical Study in Data Science Courses." *arXiv.org*, 2025.

Ze-Min Liu, Gwo-Jen Hwang, Chuang Chen, Xiang-Dong Chen, and Xindong Ye. "Integrating Large Language Models into EFL Writing Instruction: Effects on Performance, Self-Regulated Learning Strategies, and Motivation." *Computer Assisted Language Learning*, 2024.