# Does AI in the classroom facilitate deep learning in students?

August 1 2024, by Laura Grove



Students in the study found the LLM-powered AI assistant helpful for simple tasks, but sought out human teaching assistants for assignments that required critical thinking skills. Credit: Emmaline Nelsen/William & Mary

Multitasking robots collaborate with humans in large warehouses, and

chatbots respond to queries on banking websites. Artificial intelligence assistants even sort documents for law firms. William & Mary Assistant Professor of Computer Science Janice Zhang says that it's only a matter of time before AI and large language model (LLM) tools are widely used in education, as well.

To gain a better understanding of how AI assistants might fit into the classroom, Zhang worked with a team of researchers, including W&M Ph.D. students Wenhan Lyu and Yimeng (Yvonne) Wang, to conduct a longitudinal study that explored how the use of an LLM-powered AI teaching assistant affected student-learning outcomes in an entry-level computer programming course.

The study was published on July 15 in the *Proceedings of the Eleventh ACM Conference on Learning @ Scale*.

Zhang explained the importance of conducting a longitudinal study over a full semester. "Rather than performing a short-term, one-session study, we really wanted to understand how students use these kinds of tools in their natural environment," she said. "We wanted to look at longer-term impacts over time."

The researchers designed CodeTutor, an LLM-powered learning assistant that collected and stored data during interactions with students. They then divided an entry-level computer programming class into two groups with similar levels of proficiency. Human teaching assistants were available for the entire class, but students in one group were also given access to CodeTutor, which they were encouraged to use throughout the course.

At the end of the semester, students in the CodeTutor group showed statistically significant improvement in their overall course scores compared to students in the group who did not have access to the tool.

Additionally, within the CodeTutor group, students who had never used AI tools like ChatGPT improved significantly more than those who had more experience with LLM-powered AI.

On the surface, these results seem to indicate that AI assistants can help students improve grades and gain knowledge, but Lyu pointed out that other facets of the study provide greater context.

"I don't think the improvement of grades between the groups is the most interesting finding in our research," he said. "I believe our results that show how students used CodeTutor and how they interacted with it are actually the most important."

One trend that intrigued researchers was that students perceived that CodeTutors's ability to understand their questions decreased over time.

For example, students felt that CodeTutor provided helpful feedback for simpler tasks like debugging and syntax comprehension, but they expressed dissatisfaction with the tool in more advanced assignments that required [critical thinking](#). As a result, students increasingly sought out human teaching assistants toward the end of the semester.

Wang explained that part of the reason for CodeTutor's unsatisfactory responses to questions that required critical thinking is that many students asked unstructured questions. Sometimes, they would simply copy and paste questions from their assignments into CodeTutor without adding anything about their personal understanding of the question.

"Some students really wanted to finish the assignment instead of developing a deep understanding of how to solve the problem and why the professor chose that assignment," she said.

Students sometimes became frustrated and repeatedly asked CodeTutor

why a part of a problem was wrong without providing necessary context. Thus, they continued to receive unsatisfactory answers.

Wang explained that prompt quality is the key to a productive experience with AI teaching assistants. If students ask focused questions with clear goals and complete information, she said, they receive better responses.

The research team analyzed all student-generated prompts from the semester and found 63% of them to be unsatisfactory.

"I feel like this is a very critical moment and that we need to provide some instruction regarding how to actually ask good questions," Zhang said.

Learning to create successful prompts for these tools is a part of what Zhang calls generative AI literacy, defined as "the ability to effectively interact with AI tools and understand how to formulate queries and interpret responses."

She explained that it's also important for people to be able to differentiate between AI responses and human responses.

"I think there are a lot of trust issues and misinformation," said Zhang. "As technology evolves, generative AI or large model-generated responses have an increasingly human-like tone. I personally feel like there's a need to educate people and to foster generative AI literacy."

Zhang is currently collaborating with local school districts to design AI literacy camps for elementary and middle school students. The knowledge and experience provided by those camps will help students to interact effectively with AI throughout their lives, including time that they may spend in college.

Results of the study have helped to highlight gaps that need to be addressed, and Zhang emphasized the importance of teamwork in such achievements.

"The research team would like to thank all the participants for their time and engagement with our research," she wrote in an email, "including instructors of CSCI 140/141 who agreed to allow us to advertise our study, our reviewers, [funding agencies](#), and our other amazing co-authors including Dr. Yifan Sun, and Dr. Tingting (Rachel) Chuang."

Lyu points out that, although there are currently no definitive answers as to the overall effectiveness of AI learning assistants in college courses, the study produced interesting results that will drive further research.

"I'm not sure if using large language models is a disadvantage or an advantage," he said. "The students in the experimental group gained a greater improvement in the final exam, but we actually don't know if they learned those things—if they forwarded these things into their minds—or if they just used the tools for short-term results. It's a very interesting question that we want to explore in the future."

Provided by William & Mary