**Context**

As AI-based chatbots have become more widely used, students have naturally began relying on them for homework, studying, and essay writing. However, these chatbots have not been around very long, so the effects on student’s lives long term is not well understood. Many discussions around the use of AI focus on increasing worker productivity and how to keep students from cheating with AI, rather than analyzing the impact that AI use has on learning new skills.

This paper aims to quantify the impact that use of AI has on students' performance. It does so by creating two distinct chatbots: one modeled after GPT-4 and another designed specifically to promote active learning. Student performance is then evaluated both with and without access to these chatbots.

While AI tools like GPT-4 have shown significant potential to enhance productivity and provide knowledge, they also pose risks, such as inhibiting learning through overreliance or reducing the development of foundational skills. Understanding this tradeoff is essential for ensuring that AI technologies are used responsibly and effectively, particularly in contexts like education where long-term skill development is crucial.

There has been an increase in students using AI, and current lesson plans were not designed with that in mind. It would be difficult, at this point, to fully eliminate the use of AI by students. For that reason, it is important to understand the effect that it has on learning. This research represents a first step toward integrating AI into education effectively. For example, if the chatbot designed to promote learning turns out to be beneficial, or at the very least not harmful, it could be a replacement for student use. Or, if we understand the impact on learning, teachers may be able to re-design lesson plans to let students use AI in a way that still promotes learning and adoption of new skills.

**Executive Summary**

The paper found that the base GPT model improved student performance on assisted tasks but lowered it on unassisted tasks, suggesting it hinders learning of new concepts. In contrast, the tutor GPT model boosted performance on assisted tasks without significantly affecting unassisted ones, indicating that purpose-built AI models may be a valuable tool for supporting students.

**Methodology Review**

The study examined the educational impact of two AI chatbots—GPT Base and GPT Tutor—on student learning outcomes. GPT Base functioned as a standard GPT-4 assistant, delivering typical answers to prompts and acting as a straightforward tutor. GPT Tutor, on the other hand, was designed with additional features to support deeper learning. It provided multiple correct answers, incorporated teacher feedback to address common mistakes, and offered hints rather than full answers to encourage active problem-solving.

The research design was a randomized controlled trial involving nearly 1,000 students from 50 classrooms in grades 9 through 11 at a large high school in Turkey. Randomization was performed at the classroom level because students were already randomly assigned to these groups, and honors classrooms were excluded because they are not randomly assigned. The study spanned four 90-minute sessions, each with a sequence of 3 activities. First, teachers reviewed the topic with students. Second, students participated in a randomized, assisted AI session, where they either used GPT Base, GPT Tutor, or relied on textbooks and notes (control group). This session was scored to assess performance. Finally, students completed an unassisted evaluation, also scored, to measure how well they had learned the new information.

Data was collected in three main ways. At the start of the study, students completed a survey capturing their demographics and educational background. During the sessions, student performance was recorded for both the assisted practice and the unassisted evaluations. Additionally, students who interacted with AI chatbots had their chat data logged, and surveys captured their experiences using the tools.

To evaluate the impact of the interventions, the authors used a regression model to analyze student outcomes. The dependent variable, Outcome(j) , represented the normalized grade of a student in either the assisted (j = 0) or unassisted (j = 1) sessions, scaled from 0 to 1.

The independent variables GPTBasec and GPTTutorc indicate the treatment group for each class. The model controlled for prior student performance using normalized GPA from the previous year, PrevGPAi, and included fixed effects for session, grade, year, and time-related variations (θs, δg, αy, λt). Errors were clustered at the classroom level to account for correlations within groups.

Results indicate that use of the chatbots increased performance on the assisted assessment, with GPT Base improving scores by .137 (out of 1) and GPT Tutor improving scores by .361 (out of 1) relative to the control group. On the unassisted assessment, GPT Base decreased performance by .054 (out of one) relative to the control group (17% decrease). GPT Tutor’s impact on the unassisted portion was statistically significant at -.004.

The researchers also considered alternative analyses, such as examining results at the problem level. However, for this replication, we will focus on the overall differences between the GPT Base and GPT Tutor models without exploring problem-level details.

**Data Analysis**

Dataset creation:

To construct our dataset, we reverse-engineered the score and classroom data. We modeled 50 classrooms with 20 students each, creating one row per student. Normalized previous GPA values were generated using a normal distribution with a mean of 0.82 and a standard deviation of 0.11, bounded between 0 and 1.

Classrooms were randomly assigned to GPT Base, GPT Tutor, or control groups. Assignment probabilities were based on the proportion of students in each group as reported in the original study.

Fixed effects were assigned randomly as integers, including session numbers (1–4), grade levels (9–12), teachers (1–20), and graders (1–10).

Student scores were then simulated. Since standard errors were calculated at the classroom level, we generated coefficients for each classroom by using the coefficients and standard errors provided in the paper. Coefficients were drawn from a normal distribution using the reported means and standard deviations, then applied to the paper's linear regression model with a small amount of noise to calculate scores for each student.

Fixed effects were excluded from score generation, as their variance was already accounted for in the original study's regression model. This approach allowed us to focus solely on replicating the treatment effects described in the paper.

Replication setup:

To replicate the study's results, we used the simulated dataset created as described in the Data Analysis section. Our replication focused on reproducing the main regression model used in the original paper:

*\*\*Needs check\*\**

Outcome(j)ic = β0 + β1 \* GPTBasec + β2 \* GPTTutorc + β3 \* PrevGPAi + θs + δg + αy + λt + εic

*\*\*Needs check\*\**

Where:

- Outcome(j)ic is the normalized grade (0-1) for student i in class c for assisted (j=0) or unassisted (j=1) sessions

- GPTBasec and GPTTutorc are binary indicators for the treatment group of class c

- PrevGPAi is the normalized previous year GPA for student i

- θs, δg, αy, and λt are fixed effects for session, grade, year, and time-related variations

- εic is the error term

Standard errors were clustered at the classroom level to account for within-group correlations, as done in the original study.

Results of replication:

Our replication yielded results that closely aligned with the original study, with Assisted assessments: GPT Base: Coefficient = 0.135 (SE = 0.021) and GPT Tutor: Coefficient = 0.358 (SE = 0.023). And unassisted assessment of GPT Base: Coefficient = -0.052 (SE = 0.019), and GPT Tutor: Coefficient = -0.005 (SE = 0.020).

These results are very close to the original study's findings, where GPT Base improved assisted scores by 0.137 and decreased unassisted scores by 0.054, and GPT Tutor improved assisted scores by 0.361 and had a minimal effect (-0.004) on unassisted scores.

Our replication successfully reproduced the key findings: GPT Base significantly improved performance on assisted tasks but hindered performance on unassisted tasks, while GPT Tutor substantially improved assisted performance without significantly impacting unassisted performance.

**Conclusions**

Our replication study corroborates the original paper's findings, providing strong evidence for the differential impacts of AI chatbots on student learning:

1. Immediate vs. Long-term Effects: Both AI models improved immediate performance (assisted tasks), but their long-term impacts differed significantly. This highlights the importance of distinguishing between short-term assistance and actual learning.
2. Design Matters: The stark contrast between GPT Base and GPT Tutor's effects on unassisted performance underscores the critical role of AI design in educational contexts. Purpose-built AI tutors like GPT Tutor can enhance learning without the detrimental effects seen with general-purpose AI.
3. Potential for AI in Education: GPT Tutor's ability to improve assisted performance without significantly impacting unassisted performance suggests that carefully designed AI tools can be valuable educational assets. This opens up possibilities for integrating AI into curricula in ways that support rather than hinder learning.
4. Need for Caution: The negative impact of GPT Base on unassisted performance serves as a warning against uncritical adoption of general-purpose AI in educational settings. It emphasizes the need for careful consideration and possibly regulation of AI use in education.
5. Future Research Directions: This study paves the way for further research into optimizing AI for educational purposes, understanding the long-term impacts of AI use on skill development, and developing best practices for AI integration in various educational contexts.

In conclusion, our replication reinforces the original study's message: while AI has the potential to revolutionize education, its implementation must be thoughtful and purposeful to ensure it enhances rather than hinders genuine learning and skill development.

**Citation**

Bastani, Hamsa and Bastani, Osbert and Sungu, Alp and Ge, Haosen and Kabakcı, Özge and Mariman, Rei, Generative AI Can Harm Learning (July 15, 2024). The Wharton School Research Paper, Available at SSRN: https://ssrn.com/abstract=4895486 or http://dx.doi.org/10.2139/ssrn.4895486