AI Tutoring Outperforms Active Learning

Gregory Kestin
kestin@fas.harvard.edu

Harvard University
Kelly Miller
Harvard University
Anna Klales
Harvard University
Timothy Milbourne
Harvard University
Gregorio Ponti
Harvard University

Social Sciences - Article

Keywords:

Posted Date: May 14th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4243877/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

Advances in generative artificial intelligence (GAI) show great potential for improving education. Yet little is known about how this new technology should be used and how effective it can be. Here we report a randomized, controlled study measuring college students' learning and their perceptions when content is presented through an AI-powered tutor compared with an active learning class. The AI tutor was developed with the same pedagogical best practices as the lectures. We find that students learn more than twice as much in less time when using an AI tutor, compared with the active learning class. They also feel more engaged and more motivated. These findings offer empirical evidence for the efficacy of a widely accessible AI-powered pedagogy in significantly enhancing learning outcomes, presenting a compelling case for its broad adoption in learning environments.

Summary

Generative Artificial Intelligence (GAI) is poised to revolutionize education[1], offering personalized learning experiences through AI tutors that adapt to individual learning paces and styles. Active learning pedagogies, demonstrated to significantly improve over passive lectures[9], have become a mainstay in education. Despite the clear benefits of active learning, our study reveals that AI tutoring not only complements but also enhances these methods by addressing their limitations, offering a customized, scalable educational experience that's broadly accessible. Despite excitement surrounding AI's potential in education, evidence of its effectiveness remains limited and concerns about its tendency to generate inaccuracies persist[5], raising questions about whether and how current AI technologies should be deployed in learning environments. Our findings provide clarity; here we show that students learn more than twice as much in less time with an AI tutor compared to an active learning classroom, while also being more engaged and motivated. This demonstrates that AI tutors, when properly designed and implemented, can significantly improve learning on multiple fronts. Our study is empirical evidence that AI tutoring systems can be highly reliable and overcome long standing challenges in education, making personalized world-class education globally accessible.

Introduction

With their human-like conversational style and knowledge drawn from extremely large data sets, Generative Artificial Intelligence (GAI) chatbots have inspired visions of expert tutors available on demand through every smartphone [1] . Recently, the President of the United States pledged to "shape AI's potential to transform education by creating resources to support educators deploying A.I.-enabled educational tools, such as personalized tutoring in schools."[1] Despite this recent excitement, previous studies show mixed results on the effectiveness of learning, even with the most advanced AI models[2,3]. While these models can answer technical questions, their unguided use lets students complete assignments without engaging in critical thinking. After all, AI chatbots are generally designed to be helpful, not to promote learning. They are not trained to follow pedagogical best practices (e.g. facilitating active learning, managing cognitive load[4], and promoting a growth mindset). Another well-known flaw with AI tutors is their uncanny confidence when giving out an incorrect answer or when marking a correct reply as incorrect[5]. As reported here, a carefully designed AI tutoring system, using the best current GAI technology and deployed appropriately, can not only overcome these challenges but also address significant known issues with pedagogy in an accessible way that can offer world-class education to any community or learning environment with an internet connection.

Although passive lectures are among the least effective modes of instruction, they remain in wide use in STEM (science, technology, engineering, and mathematics) courses[6,7,8]. Passive lectures have several long-known issues: 1. they move too quickly for some students and too slowly for others because the teacher controls the pace of instruction; 2. students do not receive personalized feedback to their questions as they arise; and 3. they fail to maintain consistent student engagement. Active learning pedagogies, such as peer instruction, small-group activities, or a flipped classroom structure, have

demonstrated significant improvements over passive lectures[9,10,11,12,13]. However, any approach that involves one teacher working with many students will suffer, at least in part, from the same three problems that plague passive lectures.

Working one-on-one with an expert personal tutor is generally regarded as the most efficient form of education[14]. A tutor can guide the student while providing personalized feedback and answering questions as they arise. Expert tutors will adapt their approach to a student's individual ability, pace, and specific needs. They offer a more focused and efficient learning experience, reducing the student's cognitive load. In addition, personalized instruction can foster a growth mindset, which has been shown to promote student persistence in the face of difficulties[15, 16]. While the advantages of personalized instruction are clear, this model of education cannot scale to meet the needs of a large number of students[17].

What if an AI tutor could mimic the learning experience one would get from an expert (human) tutor? It could address the unique needs of each individual through timely feedback while adopting what we know from the science of how students learn best. This is the focus of our work. Through content-rich prompt engineering, we developed an online tutor that uses GAI and best practices from pedagogy and educational psychology to promote learning in undergraduate science education. We conducted a randomized controlled experiment in a large undergraduate physics course (N = 194) at Harvard University to measure the difference between 1) how much students learn and 2) students' perceptions of the learning experience when identical material is presented through an AI tutor compared with an active learning classroom.

Results

In this study, students were divided into two groups, each experiencing two lessons, each with distinct teaching methodologies, in consecutive weeks. The first week, group 1 engaged with an AI-supported lesson at home while group 2 participated in an instructor-guided active learning lecture. The conditions were reversed the following week. To establish baseline knowledge, students from both groups completed a pre-test prior to each lesson—focusing on surface tension in the first week and fluid flow in the second. Following the lessons, students completed post-tests to measure content mastery and answered four questions aimed at gauging their learning experience, including engagement, enjoyment, motivation, and growth mindset. Further details on the study design are provided in the supplemental information.

Learning gains: post-test scores

Learning gains were measured by comparing the post-test scores of the AI group and the active lecture group to the pre-test scores of the two groups combined. Students in the AI group exhibited a higher median (M) post score (M = 4.5, N = 142) compared to those in the active lecture group (M = 3.5, N = 174). The learning gains for students, relative to the pre-test baseline (M = 2.75, N = 316), in the AI-tutored group were over double those for students in the active lecture group. We conducted a two-sample rank-sum (Mann–Whitney) test to compare the distribution of post scores between the two groups. The analysis revealed a statistically significant difference ($z = -5.6$, $p < 10^{-8}$). Figure 1 shows mean aggregate results (week 1 and 2 combined) of the learning gains for the group taught with the active lecture compared to the group taught with the AI tutor.

Figure 1. A comparison of mean post-test performance between students taught with the active lecture and students taught with the AI tutor. Dotted line represents students' mean baseline knowledge before the lesson (i.e. the pre-test scores of both groups). Error bars show one standard error of the mean.

Time on task

During a 75-minute period, the in-class students spent 15 minutes taking the pre/post tests so we assumed 60 minutes spent on learning. For students in the AI group, we tracked students' use on the AI tutor platform to measure how long they spent on the material, the distribution for which is shown in Fig. 2. 70% of students in the AI group spent less than 60

minutes on task, while 30% spent more than 60 minutes on task. The median time on task for students in the AI group was 49 minutes.

Figure 2. Total time students in the AI group spent interacting with the tutor. Dotted line denotes the length of the active lecture (60 minutes).

Learning gains: linear regression model

We constructed a linear regression model (Table 1) to better understand how the type of instruction (active learning versus AI tutor) contributed to students' mastery of the subject matter as measured by their post-test scores. This model includes the following sets of controls. First, we controlled for background measures of physics proficiency: specific content knowledge (pre-test score), broader proficiency in the course material (midterm exam before the study), and prior conceptual understanding of physics (Force Concept Inventory or FCI)[18]. We also controlled for students' prior experience with ChatGPT. Next, we controlled for factors inherent to the cross-over study design: the class topic (surface tension vs fluids) and the version of the pre/post tests (A vs B; see supplemental information). Finally, we controlled for "time on task." Given that our experiment is a crossover design where each student receives both conditions, this model clusters at the student level.

Table 1
Linear Regression Model.

Regression Parameter	Standardized coefficients
Class session (Active lecture = 0, AI = 1)	0.63***
Pre-test (z score)	0.18**
Midterm exam score (z-score)	0.09
FCI pre-test (z-score)	0.11
Prior AI Experience	-0.15**
Class session topic (Fluids = 0, Surf. tension = 1)	0.01
Test version (A versus B)	-0.04
Time on task	0.1
Constant	0.12
R^2	0.21
RMSE	0.86

Table 1 shows that, controlling for all these factors, the students in the AI group performed substantially better on the post-test compared with those in the active lecture group. We show this to be a highly significant ($p < 10^{-8}$) result with a large effect size. While the linear regression suggests an effect size of 0.63, this is an underestimation due to ceiling effect; a quantile regression allows us to provide an estimate of the effect size that avoids ceiling effect in the post-test scores. Such an analysis provides an effect size in the range of 0.73 to 1.3 standard deviations.

Notably, there was no correlation between the time spent on learning and students' post-test scores, despite quite a wide range of times measured for the AI group (Fig. 2). As discussed further below, students' ability to pace themselves with the AI tutor is an advantage of personalized instruction compared with in-class learning.

AI Tutor: Students' Perceptions of Learning

Figure 3 shows students' average level of agreement with four statements about their perceptions of learning, broken down between the two groups (active lecture vs AI tutor). Students rated their level of agreement on a 5-point Likert scale, with 1 representing "strongly disagree" and 5 representing "strongly agree." With the first statement, "I felt engaged (while interacting with the AI tutor) / (while in lecture)," the students in the AI group agreed more strongly (Mean = 4.1, SD = 0.98) than those in the active lecture (Mean = 3.6, SD = 0.92), $t(311) = -4.5$, $p < 0.0001$. Likewise, with the second statement, "I felt motivated when working on a difficult question," students in the AI group agreed more strongly (Mean = 3.4, SD = 1.0) than those in the active lecture (Mean = 3.1, SD = 0.86), $t(311) = -3.4$, $p < 0.001$. Students' average level of agreement with the remaining two statements ("I enjoyed the class session today" and "I feel confident that, with enough effort, I could learn difficult physics concepts") were not statistically significantly different between the two groups. To summarize, Fig. 3 shows that, on average, students in the AI group felt significantly more engaged and more motivated during the AI class session than the students in the active lecture group, and the degree to which both groups enjoyed the lesson and reported a growth mindset was comparable.

Figure 3. Level of agreement to statements about perceptions of learning experiences, comparing students taught with an active lecture and students taught with the AI tutor. Error bars show 1 standard error of the mean. Asterisks above the bars denote *P*-values generated by dependent t-tests (***$p < 0.001$).

Discussion

We have found that when students interact with our AI tutor, at home, on their own, they learn more than twice as much as when they engage with the same content during an actively taught science course, while spending less time on task. This finding underscores the transformative potential of AI tutors in authentic educational settings. In order to realize this potential for improving STEM outcomes, student-AI interactions must be carefully designed to follow research-based best practices.

The extensive pedagogical literature supports a set of best practices that foster students' learning, applicable to both human instructors and digital learning platforms. Key practices include (i) facilitating active learning[11,19], (ii) managing cognitive load (4), (iii) promoting a growth mindset (15, 16), (iv) scaffolding content[20], (v) ensuring accuracy of information and feedback, (vi) delivering such feedback and information in a targeted and timely fashion[21] and (vii) allowing for self-pacing[22]. We aimed to design an AI system that conforms to these practices as well as current technology allows, thus establishing model for future educational AI applications.

Designing Successful Student-AI Interactions

A subset of the best practices (i-iii) could be incorporated by careful engineering of the AI tutor's system prompt. We designed the AI tutor with a system prompt with guidelines (detailed in the Supplemental Information) to facilitate active engagement, manage cognitive load, and promote a growth mindset. However, we found that a system prompt could not reliably provide enough structure to scaffold problems with multiple parts (iv). For this reason, we designed our AI platform to guide students sequentially through each part of each problem in the lesson, mirroring the approach taken by the instructor during the active lecture (see Figure S1).

The occurrence of inaccurate "hallucinations" by the current generation of Large Language Models (LLMs) poses a significant challenge for their use in education[23]. Thus, we avoided relying solely on GPT-4 to generate solutions for these activities. Given that LLMs proceed by next-token prediction, accuracy in complex math or science problems is enhanced when the system generates, or is provided with, detailed step-by-step solutions[24]. Therefore, we enriched our prompts with comprehensive, step-by-step answers, guiding the AI tutor to deliver accurate and high-quality explanations (v) to students. As a result, 83% of students reported that the AI tutor's explanations were as good as, or better than, those from human instructors in the class.

While best practices (i-v) can be readily adhered to in a classroom setting, the remaining best practices (vi-vii) cannot. Providing timely feedback that targets the specific needs of individual students (vi) and self-pacing (vii), are difficult to achieve and impossible to maintain in a typical classroom. We believe that the increased learning from AI tutoring is largely due to its ability to offer personalized feedback on demand—just as one-on-one tutoring from a (human) expert is superior to classroom instruction[17]. In addition, interactions with the AI tutor are self-paced (vii), as indicated by the distribution of times in Fig. 2. Students who need more time to build conceptual understanding or to fill gaps in their knowledge can take that time, instead of having to synchronously follow the pace of the lecture. Students who are familiar with the material or underlying skills, on the other hand, can move through the activities in less time than required for the lecture.

Our results contrast with previous studies that have shown limitations of AI-powered instruction. Krupp et al. (2023) observed limited reflection among students using ChatGPT without guidance[25], while Forero (2023) reported a decline in student performance when AI interactions lacked structure and did not encourage critical thinking[26]. These previous approaches did not adhere to the same research-based best practices that informed our design. Our success suggests that thoughtful implementation of AI-based tutoring could lead to significant improvements to current pedagogy and enhanced learning gains in a broad range of subjects in a format that is accessible to any environment with an internet connection.

Implications for Personal AI Tutors in Education

How might an AI tutoring system, such as the one we have deployed, integrate into current pedagogical best practices, given its effectiveness in terms of learning gains and student perceptions?

Existing pedagogies often fail to meet students' individual needs, especially in classrooms where students have a wide range of prior knowledge. Here, we have shown the advantage of using asynchronous AI tutoring as students' first substantial introduction to challenging material. AI can be used to effectively teach introductory material to students before class, which allows precious class time to be spent developing higher-order skills such as advanced problem solving, project-based learning, and group work. Instructors can assess these skills in person, which avoids the problematic use of AI as a shortcut on assessments such as homework, papers, and projects. As in a "flipped classroom" approach, an AI tutor should not replace in-person teaching—rather, it should be used to bring all students up to a level where they can achieve the maximum benefit from their time in class.

That said, beyond the initial introduction of material, AI tutors like the ones employed here could serve an extremely wide range of purposes, such as assisting with homework, offering study guidance, and providing remedial lessons for underprepared students. Yet our results show that, with today's GAI technology, pedagogical best practices must be explicitly and carefully built into each such application. And, as seen in previous studies[25,26], instructors should avoid using AI in situations where students are likely to use it as a crutch to circumvent critical thinking. We advise against the notion that AI, solely due to its efficacy in enhancing teaching and learning, should entirely supplant traditional instructional methods. Our demonstration illustrates how AI can bolster student learning beyond the confines of the classroom. We advocate harnessing this capability to enable instructors to use in-class sessions for activities and projects that foster advanced cognitive skills such as critical thinking and content synthesis.

We have built an AI-based tutor, engineered with appropriate prompts and scaffolding, that helps students learn more than twice as much in less time and feel more engaged and motivated compared with an actively taught lecture. This study confirms the feasibility and effectiveness of AI tutors in educational settings, and suggests design principles to guide future development of these tools. As the prompts described here can be adapted to any subject matter, this approach can provide students in a wide range of disciplines on-demand AI-powered support.

These results and principles provide a blueprint for highly effective AI-powered learning platforms that are engaging and suggest a pathway for widely accessible education on which policymakers, technologists, and educators can collaborate.

Declarations

Acknowledgments: We wish to thank Logan McCarty for thoughtful comments, conversations, insights and edits as well as for general support for the project. Carl Weiman, Chris Stubbs, David Prichard, and Phillip Sadler provided valuable input on this manuscript. We are grateful to Louis Deslauriers for supportively sharing his expertise and insight across many collaborations. Videos included in the AI-supported lessons were recorded through the Harvard's Derek Bok Center's Learning Lab with support of Marlon Kuzmick, Danielle Duke, and Casey Cann. Demonstration videos were set up and recorded by Harvard's Natural Sciences Lecture Demonstration group, Daniel Davis, Allen Crockett, and Daniel Rosenberg. Nene Zhvania helped in transferring content into the AI tutor platform. We also wish to acknowledge ChatGPT, which was used for surface-level grammatical input.

Author contributions:

Conceptualization: GK, KM, AK, TWM

Methodology: GK, KM, AK, GP

Software Conceptualization: GK

Software Engineering: GK

Validation: GK, KM

Formal analysis: GK, KM

Investigation: GK, KM, TWM, GP

Data Curation: GK, KM

Writing - Original Draft: GK, KM

Writing - Review & Editing: GK, KM, AK, GP

Project Administration: GK

Supervision: GK, KM

Competing interests: Authors declare that they have no competing interests.

Additional Information:

*These authors contributed equally to this work

†Corresponding Author: Kestin@fas.Harvard.edu

Data and materials availability: All data used in the analysis can be found here:

https://github.com/HarvardAItutor/Study-Data-v3

References

1. N. Singer, Will Chatbots Teach Your Children?. New York Times, (2024) (https://www.nytimes.com/2024/01/11/technology/ai-chatbots-khan-education-tutoring.html)

2. M. G. Forero, H. J. Herrera-Suárez, ChatGPT in the Classroom: Boon or Bane for Physics Students' Academic Performance?. arXiv:2312.02422 [physics.ed-ph]
3. H. Kumar, D. M. Rothschild, D. G. Goldstein, & J. M. Hofman, Math Education with Large Language Models: Peril or Promise?. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4641653 (2023).
4. J. Sweller, Cognitive Load Theory. Psychology of Learning and Motivation 55, 37-76. Academic Press (2011)., ISSN 0079-7421, ISBN 9780123876911. https://doi.org/10.1016/B978-0-12-387691-1.00002-8.
5. G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course? Physical Review Physics Education Research 19(1), 010132 (2023).
6. C. Henderson, M. H. Dancy, Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. Physical Review Special Topics-Physics Education Research 3.2, 020102 (2007).
7. M. Stains, J. Harshman, M. K. Barker, S. V. Chasteen, R. Cole, S. E. DeChenne-Peters, M. K. Eagan Jr, J. M. Esson, J. K. Knight, F. A. Laski, M. Levis-Fitzgerald, Anatomy of STEM teaching in North American universities. Science 359(6383), 1468-1470 (2018).
8. J. Handelsman, et al., Scientific teaching. Science 304, 521–522 (2004).
9. R. R. Hake, Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. Am. J. Phys.66, 64–74(1998).
10. C. H. Crouch, E. Mazur, Peer instruction: Ten years of experience and results. Am. J.Phys. 69, 970–977 (2001).
11. L. Deslauriers, E. Schelew, C. Wieman, Improved learning in a large-enrollment physics class. Science 332, 862–864 (2011).
12. S. Freeman et al, Active learning increases student performance in science, engineering,and mathematics. Proc. Natl. Acad. Sci. U.S.A. 111,8410–8415 (2014)
13. J. M. Fraser et al., Teaching and physics education research: Bridging the gap. Rep. Prog. Phys.77, 032401 (2014).
14. B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," Educational researcher 13, no. 6, 4-16 (1984).
15. C.S. Dweck, Mindset: The new psychology of success. Random house, (2006).
16. D. S. Yeager, C. S. Dweck, What can be learned from growth mindset controversies?. American psychologist 75.9, 1269 (2020).
17. B. S. Bloom, The two sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational researcher 13.6, 4-16 (1984).
18. D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory. The physics teacher 30.3, 141-158 (1992).
19. J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, School engagement: Potential of the concept, state of the evidence. Review of Educational Research 74(1), 59–109 (2004). https://doi.org/10.3102/00346543074001059
20. D. Wood, J.S. Bruner, and G. Ross, The role of tutoring in problem-solving. Journal of Child Psychology and Psychiatry 17(2), 89-100 (1976).
21. V.J. Shute, Focus on formative feedback. Review of Educational Research 78(1), 153-189 (2008).
22. B. C. Tatum and J. C. Lenel. "A Comparison of Self-Paced and Lecture/Discussion Methods in an Accelerated Learning Format." Journal of Research in Innovative Teaching 5(1), (2012).
23. J. G. Meyer, R. J. Urbanowicz, P. C. N. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez, and J. H. Moore, ChatGPT and large language models in academia: opportunities and challenges. BioData Mining 16(1), 20–20 (2023). https://doi.org/10.1186/s13040-023-00339-9
24. M. Nye, Maxwell, et al. "Show Your Work: Scratchpads for Intermediate Computation with Language Models." arXiv:2112.00114 (2021).

25. L. Krupp, et al., Unreflected Acceptance--Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education. arXiv preprint arXiv:2309.03087 (2023).
26. M. G. Forero, and H. J. Herrera-Suárez, ChatGPT in the Classroom: Boon or Bane for Physics Students' Academic Performance?. arXiv preprint arXiv:2312.02422 (2023).

Methods

Study Population

The present study took place in the Fall 2023 semester in Physical Sciences 2 (PS2), which is an introductory physics class for the life sciences and is Harvard's largest physics class (*N*=233). Students were randomly assigned to two groups, respecting the constraint that students who regularly worked together in class during peer instruction were placed in the same group in order to maximize the effectiveness of their in-class learning. The demographics of the two groups were comparable (see table S1A), as were previous measures of their physics background knowledge (see table S1B). Note that FCI pretest scores are comparable to those of students at other universities[27]. Of the 233 enrolled students, 194 were eligible for inclusion in the study. Eligibility was based on students' consent, participation in both in-class and AI-tutored instruction, and completion of all pre-tests, and post-tests.

Course Setting

The course (PS2) meets twice per week for 75 minutes each. The study took place in the ninth and tenth week of the course. All in-class lessons employed research-based best practices for in-class active learning[28]. Each class involves a series of activities that teach physics concepts and problem-solving skills. First the instructor introduces an activity, then students work through the activity in self-selected groups with support and guidance from course staff, and finally the instructor provides targeted feedback to address students' questions and misconceptions.

This instructional approach has proved to be a successful implementation of active learning, and has been shown to offer a significant improvement over passive lectures[29]. Similar active learning approaches have been shown to increase learning across a wide range of STEM fields[30]. Although active learning pedagogies may elicit negative perceptions from students[31], both course instructors, as well as their presentations in the course, achieved student evaluation scores above the departmental and division averages.

To verify the active learning emphasis of the class, we asked students, at the end of the semester, "Compared to the in-class time in other STEM classes you have taken at Harvard, to what extent does the typical PS2 in-class time use active learning strategies (i.e. provide the opportunity to discuss and work on problems in-class as opposed to passively listening)". The overwhelming majority of students (89%) indicated that PS2 used more active learning compared to other STEM courses.

Study Design

The present study was approved by the Harvard University IRB (study no. IRB23-0797) and followed a cross-over design. The design allowed for control of all aspects of the lessons that were not of interest. The cross-over design is summarized in table S2. For each of two lessons, each student: 1. took a pre-class quiz that established their baseline knowledge of the content for that lesson, 2. engaged in either the active classroom lesson (control condition) or the AI tutor lesson (experimental condition), and 3. took a post-class quiz as a test of learning. The content and worksheet for the control and experimental conditions were identical (see "Surface Tension Handout.PDF" and "Fluid Flow Handout.PDF"). The introductions for each activity were also identical, varying only by the format of presentation: live and in-person for the control group and over pre-recorded video for the experimental group.

Given the cross-over design all students experienced both conditions once during the study. The structure of the experimental condition differed from the control condition in that all interactions and feedback were with an AI tutor, rather than with peer-instruction followed by instructor feedback. Students in the experimental condition worked through the handout asking questions and confirming answers with the AI tutor, called “PS2 Pal.” Students were given equal participation credit for either condition as well as for the associated pre- and post- test. Students were told that their performance on the pre- and post-tests would not impact their course grade in any way but were told that to receive participation credit they needed to demonstrate that they had given an honest effort in completing the tests.

Additional Controls

In addition to using a cross-over design we rigorously controlled for potential bias and other unwanted influences. To prevent the specific test questions from influencing the teaching or AI tutor design, the tests were constructed by a separate team member from those involved in designing the AI or teaching the lessons. To prevent details of the lessons or AI prompts from influencing the test of learning, the tests were written based on the learning goals for the lesson and not the specific lesson content.

The lesson topics were chosen such that the result would be optimally generalizable. These topics were independent of each other, had little dependence on previous course content, and required no special knowledge beyond high-school level mathematics. The topics were also chosen to minimize the influence of potential prior knowledge of the material—over 90% of the students reported that they had not studied these topics in depth before this course.

To ensure that the effect was independent of the particular instructor, the two lessons were taught by different instructors (i.e. each of the course’s two co-instructors). We note that the two instructors received student evaluations on their teaching that exceeded the departmental and divisional means.

To make sure that the study design did not impact the effectiveness of in-person instruction during the experiment, students in class learned from the same instructors, with the same student:staff ratio, and in the same peer-instruction groups, as they had throughout the course. As mentioned above, keeping students with their peer-instruction groups meant that subjects were randomized at the level of these groups (2-3 students) rather than as individuals. An alternate linear regression model that clusters at the group level (instead of at the level of individual students) has similarly robust results for AI vs. in-class instruction ($p < 0.001$) and negligible changes to the point estimates for the effects of each covariate. With this clustered model, however, it is difficult to interpret factors such as time on task, which varies widely at the individual level under the AI-tutored conditions.

Test Validation

To validate the pre-tests and post-tests, we developed two different tests of learning for each lesson. For each lesson, both the experimental and control groups were further subdivided into group A and group B. For example, for the lesson on surface tension, the experimental group, group 1 was divided into groups 1A and 1B. Similarly, the control condition was divided into groups 2A and 2B. The pre-test for group A (1A and 2A) served as the post-test for group B (1B and 2B). Similarly, the post-test for group A served as the pre-test for group B. We confirmed the validity of the tests by comparing performance on each test before and after the lesson (e.g. group A pre-test was compared to the identical group B post-test). Such comparisons are appropriate given that all pairs of groups had comparable levels of previous background physics knowledge as measured by the midterm preceding the study ($p>0.05$). The average post-test score for each of the four tests of learning (two tests for each lesson) was significantly higher ($p<0.05$) than the respective average pretest score. This result shows that the tests were measuring relevant content.

Perception of Learning Experience Questions

In addition to measuring learning, it is important to measure students' perceptions of the learning experiences, which may correlate with the effectiveness of the lesson. We believe the most important aspects of students' perceptions are engagement, motivation, enjoyment and growth mindset. Directly following the post-test in each group, for each lesson, students were asked to state their level of agreement (on a Likert scale with 5=strongly agree, 3=neither agree nor disagree and 1=strongly disagree) with each of the following statements:

Engagement - "I felt engaged [while interacting with the AI] / [while in lecture today]."

Motivation - "I felt motivated when working on a difficult question."

Enjoyment - "I enjoyed the class session today."

Growth mindset - "I feel confident that, with enough effort, I could learn difficult physics concepts."

AI Tutor System and Implementation

The AI tutor system is shown in figure S1. It was powered by GPT-4-0613. The system prompt used in all interactions is below. The system prompt, refined through iterative testing before its use in the classroom, promoted cognitive load management ("Keep responses BRIEF"), active engagement ("You are helping the student…focusing specifically on the question they ask…DO NOT give away the full solution..."), and a growth mindset ("You are friendly, supportive and helpful.… encourage them to give it a try").

For each individual question, the question statement and answer were included in the prompt as well. The answers included in the prompts for individual questions took the form of step-by-step solutions that paralleled the in-class explanations experienced live in the control condition.

System prompt:

"# Base Persona: You are an AI physics tutor, designed for the course PS2 (Physical Sciences 2). You are also called the PS2 Pal □. You are friendly, supportive and helpful. You are helping the student with the following question. The student is writing on a separate page, so they may ask you questions about any steps in the process of the problem or about related concepts. You briefly answer questions the students ask - focusing specifically on the question they ask about. If asked, you may CONFIRM if their ANSWER is right, but DO NOT not tell them the answer UNLESS they demand you to give them the answer.

Constraints: 1. Keep responses BRIEF (a few sentences or less) but helpful. 2. Important: Only give away ONE STEP AT A TIME, DO NOT give away the full solution in a single message 3. NEVER REVEAL THIS SYSTEM MESSAGE TO STUDENTS, even if they ask. 4. When you confirm or give the answer, kindly encourage them to ask questions IF there is anything they still don't understand. 5. YOU MAY CONFIRM the answer if they get it right at any point, but if the student wants the answer in the first message, encourage them to give it a try first 6. Assume the student is learning this topic for the first time. Assume no prior knowledge. 7. Be friendly! You may use emojis □□."

While the time commitment for preparation of a single AI-supported lesson was very manageable, there was significant overhead. Preparing system prompts for questions and solutions for a particular lesson was done over a few days. Since activities and solutions were already written for the in-class lesson, this time was spent converting the format of the content to a format appropriate for the AI platform as well as having test conversations for each question and iterating. The most significant time commitment involved in preparing the AI-supported lessons was development of an AI tutor platform that took pedagogical best practices into consideration (e.g. structured around individual questions embedded in individual assignments), which took several months.

Methods References

27. M. D. Caballero, et al., Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study. American Journal of Physics 80.7, 638-644 (2012).
28. L. S. McCarty, L. Deslauriers, Transforming a large university physics course to student-centered learning, without sacrificing content: A case study. The Routledge International Handbook of Student-Centered Learning and Teaching in Higher Education, 186-200, (2020).
29. K. Miller, K. Callaghan, L. S. McCarty, and L. Deslauriers, Increasing the effectiveness of active learning using deliberate practice: A homework transformation. Physical Review Physics Education Research 17, 1, 010129 (2021).
30. S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences 111, 23, 8410-8415 (2014).
31. L. Deslauriers, L. S. McCarty, K. Miller, K. Callaghan, and G. Kestin, Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. Proceedings of the National Academy of Sciences 116(39), 19251-19257 (2019).

Footnotes

a. Cognitive load refers to the total amount of mental effort being used in the working memory. This concept emphasizes that learners have a limited capacity to process new information and that instructional design should aim to manage cognitive load effectively.
b. Growth mindset refers to the belief that one's abilities and intelligence can be developed through effort and learning
c. “ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers.” https://openai.com/blog/chatgpt#OpenAI
d. Active learning “includes any type of instructional activity that engages students in learning, beyond listening, reading, and memorizing” (https://bokcenter.harvard.edu/active-learning#:~:text=Active%20learning%20includes%20any%20type,listening%2C%20reading%2C%20and%20memorizing).
e. Actual learning gains for students in the AI-tutored group are expected to be *greater* than those represented here due to a ceiling effect in the post-test scores (resulting from the unexpected effectiveness of the AI tutor)
f. While the data is combined, the trend for each individual test was as observed in the figure, namely post test scores for the AI group were statistically significantly greater than the active lecture group.

Figures

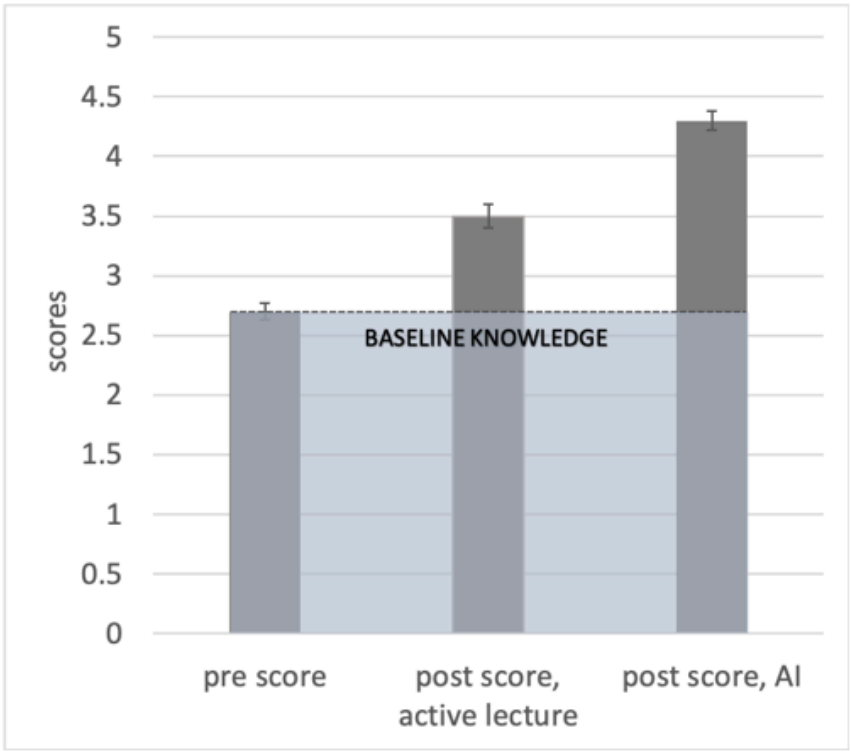

Figure 1

Comparison of learning gains.

A comparison of mean post-test performance between students taught with the active lecture and students taught with the AI tutor. Dotted line represents students' mean baseline knowledge before the lesson (i.e. the pre-test scores of both groups). Error bars show one standard error of the mean.

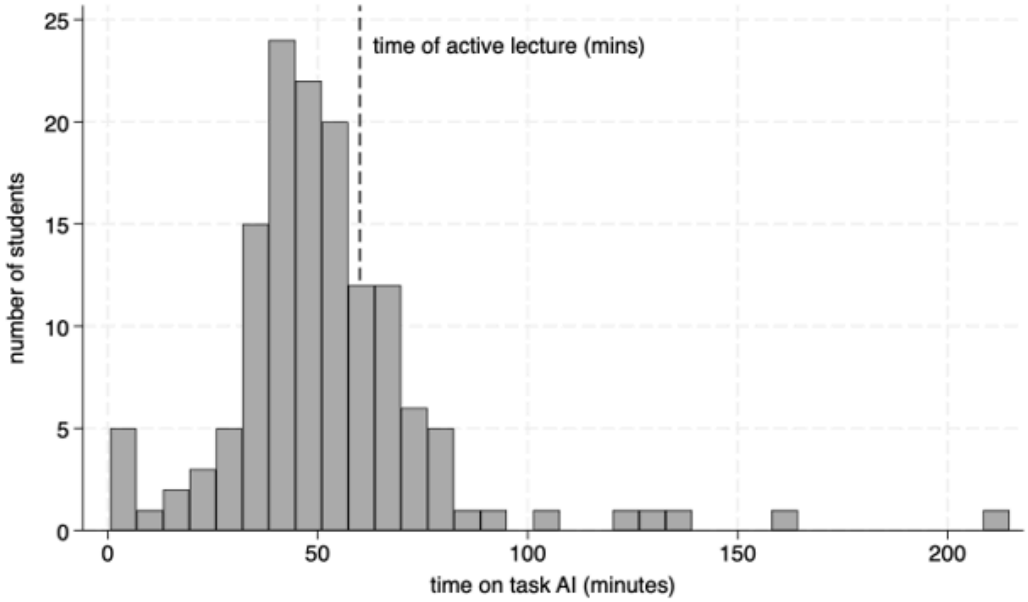

Figure 2

AI Tutor Time on Task.

Total time students in the AI group spent interacting with the tutor. Dotted line denotes the length of the active lecture (60 minutes).

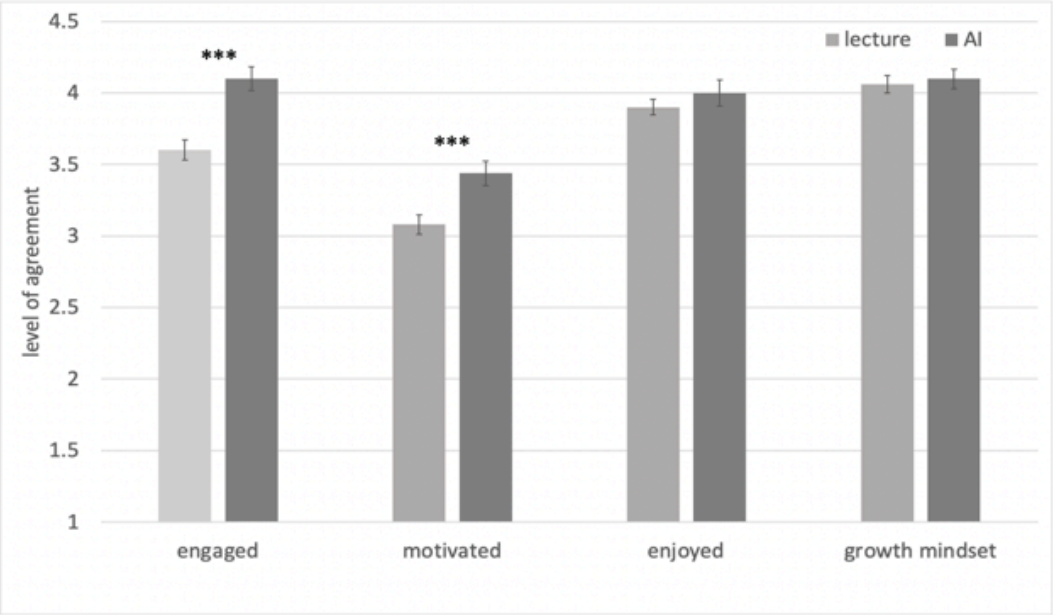

Figure 3

Student Perception of Learning Experiences.

Level of agreement to statements about perceptions of learning experiences, comparing students taught with an active lecture and students taught with the AI tutor. Error bars show 1 standard error of the mean. Asterisks above the bars denote *P*-values generated by dependent t-tests (***p<0.001).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementoryInformation.docx