scientific reports

Check for updates

OPEN

AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting

Greg Kestin[1,3]✉, Kelly Miller[2,3], Anna Klales[1], Timothy Milbourne[1] & Gregorio Ponti[1]

Advances in generative artificial intelligence show great potential for improving education. Yet little is known about how this new technology should be used and how effective it can be compared to current best practices. Here we report a randomized, controlled trial measuring college students' learning and their perceptions when content is presented through an AI-powered tutor compared with an active learning class. The novel design of the custom AI tutor is informed by the same pedagogical best practices as employed in the in-class lessons. We find that students learn significantly more in less time when using the AI tutor, compared with the in-class active learning. 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.

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]. The President of the United States, at the time of this investigation in 2023, pledged to "shape AI's potential to transform education by creating resources to support educators deploying AI-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],[1] and promoting a growth mindset).[2] Another well-known flaw of AI tutors is their uncanny confidence when giving an incorrect answer or when marking a correct reply as incorrect[5].[3] 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 pedagogical challenges 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 science, technology, engineering, and mathematics (STEM) courses[6–8]. 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

[1] Cognitive load refers to the total amount of mental effort 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.
[2] Growth mindset refers to the belief that one's abilities and intelligence can be developed through effort and learning.
[3] "ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers." https://openai.com/blog/chatgpt#OpenAI.

[1]Department of Physics, Harvard University, 17 Oxford Street, Cambridge, MA 02138, USA. [2]School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, MA 02138, USA. [3]Greg Kestin and Kelly Miller contributed equally to this work. ✉email: Kestin@fas.Harvard.edu

fail to maintain consistent student engagement. Active learning pedagogies,[4] such as peer instruction, small-group activities, or a flipped classroom structure, have demonstrated significant improvements over passive lectures[9–14]. 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 with an expert personal tutor is generally regarded as the most efficient form of education[15]. 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, managing the student's cognitive load. In addition, personalized instruction can foster a growth mindset, which has been shown to promote students' persistence in the face of difficulties[16,17]. While the advantages of personalized instruction are clear, this model of education cannot scale to meet the needs of a large number of students[15].

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 is known from the science of how students learn best. This is the focus of our work. Through a design that involves targeted, 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, with a student population broadly representative of those found across a range of institutions, 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. During, the first week, group 1 engaged with an AI-supported lesson at home while group 2 participated in an active learning lesson in class. 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 each lesson, 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.

Learning gains: post-test scores

Learning gains were measured by comparing the post-test scores of the AI group and the in-class active learning 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 in-class active learning group (M = 3.5, N = 174). The median learning gains for students, relative to the pre-test baseline (M = 2.75, N = 316), in the AI-tutored group were over double[5] those for students in the in-class active learning group. We conducted a two-sample rank-sum (Mann–Whitney) test to compare the distribution of post-scores of the two groups. The analysis revealed a statistically significant difference ($z = -5.6$, $p < 10^{-8}$). Figure 1 shows mean aggregate results (weeks 1 and 2 combined)[6] of the learning gains for the group taught with in-class active learning compared to the group taught with the AI tutor.

Time on task

During a 75-minute period, the in-class students spent 15 minutes taking the pre- and post-tests; we assume 60 minutes spent on learning. For students in the AI group, we tracked their use of 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.

Learning gains: linear regression model

We constructed a linear regression model (Table S1) to better understand how the type of instruction (in-class 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]. Second, we controlled for students' prior experience with ChatGPT. Third, we controlled for factors inherent to the crossover study design: the class topic (surface tension vs fluids) and the version of the pre/post tests (A vs B). Finally, we controlled for time on task. Given that our experiment is a crossover design in which each student experiences both conditions, this model clusters at the student level.

[4] 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).

[5] Actual learning gains for students in the AI-tutored group are expected to be *greater* than those represented by metrics presented here due to a ceiling effect in the post-test scores (resulting from the unexpected effectiveness of the AI tutor). Note that measures that are less sensitive to ceiling effect, such as the median, will be *more* reliable than measures that are more sensitive to ceiling effect, such as straight gain or mean.

[6] While the data is combined, the trend for each weeks' tests were as observed in the figure, namely post-test scores for the AI group were greater than the in-class active learning group's scores.

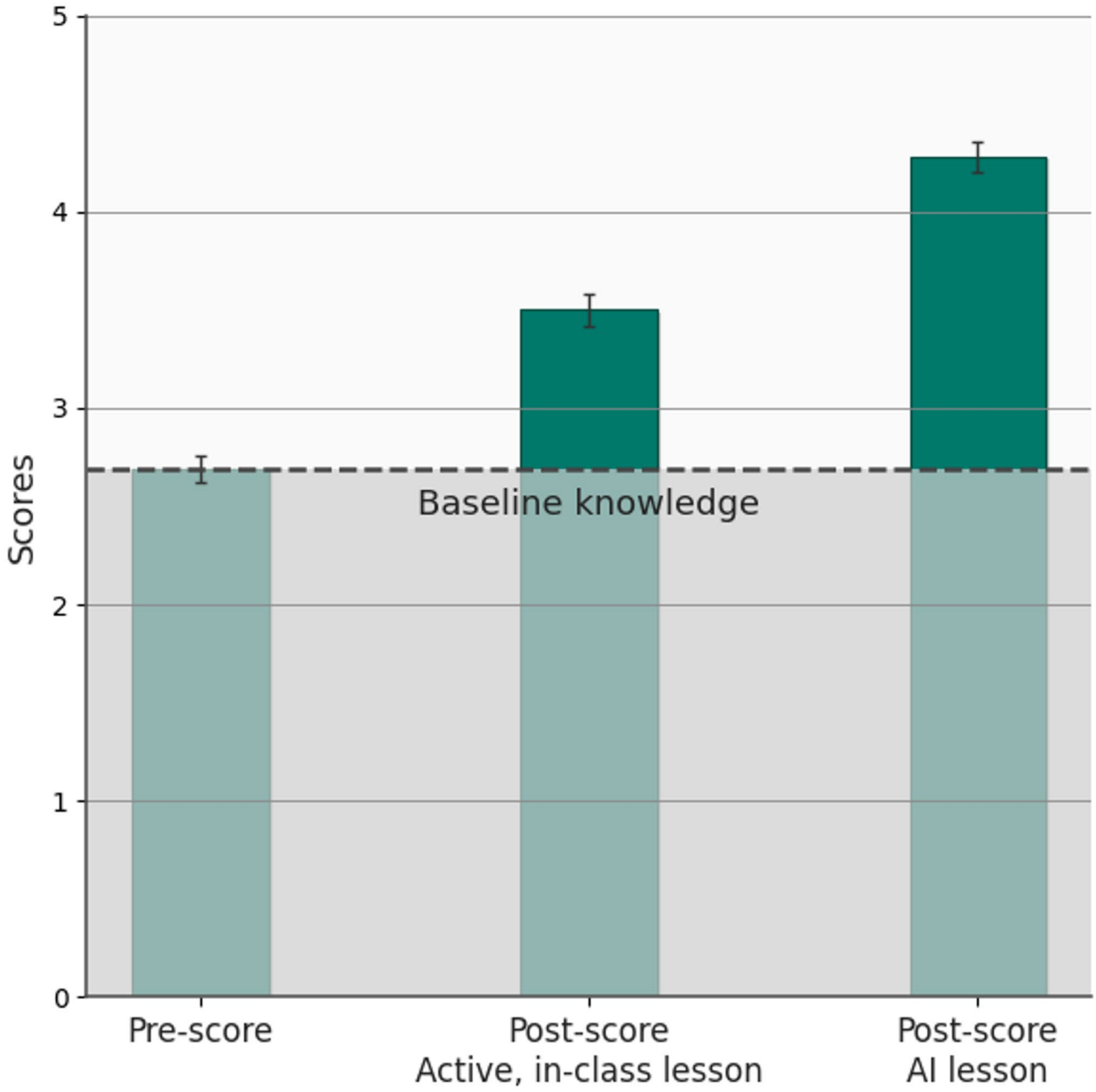

Fig. 1. Comparison of learning gains: A comparison of mean post-test performance between students taught with the in-class active learning and students taught with the AI tutor. Dotted line represents students' mean baseline knowledge before the lesson (i.e., the pre-test scores of both groups). Error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

Table S1 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 in-class active learning 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 on task 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 (in-class active learning vs. AI tutor). Students rated their level of

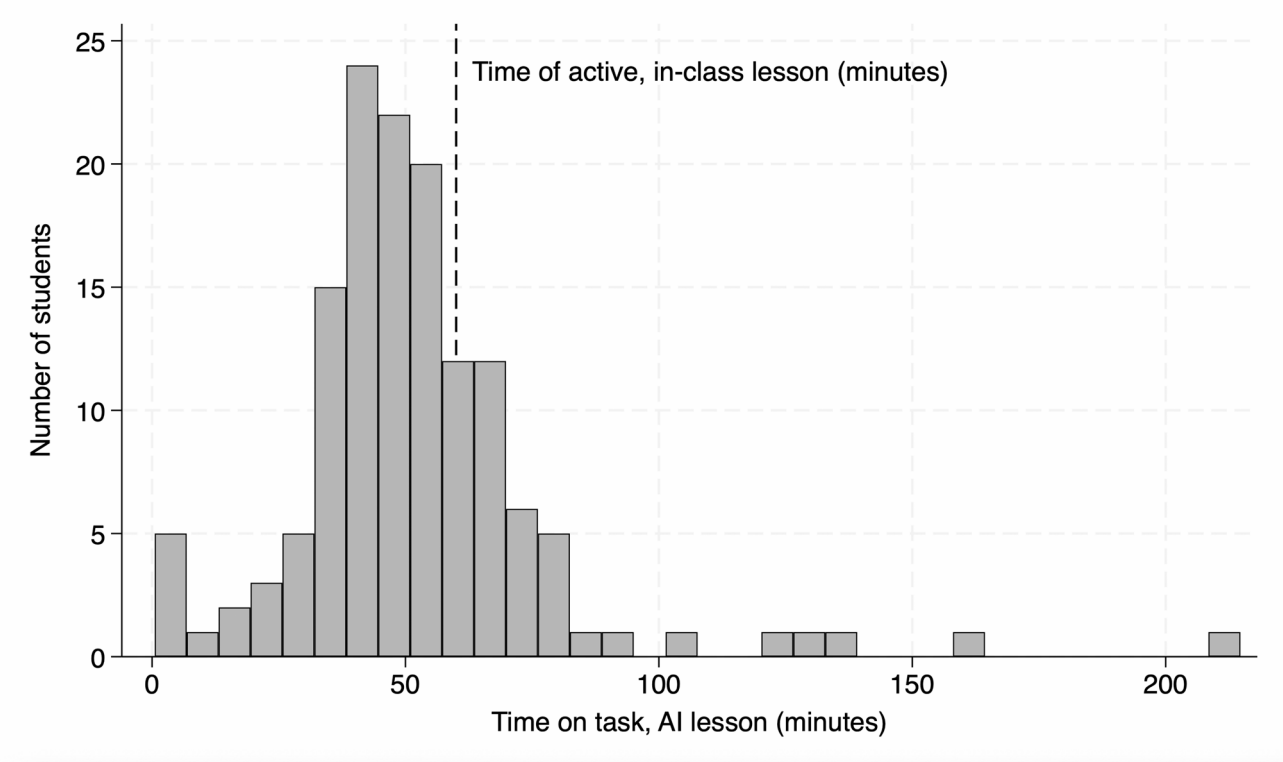

Fig. 2. AI tutor time on task: Total time students in the AI group spent interacting with the tutor. Dotted line denotes the length of the in-class active learning lesson (60 minutes).

agreement on a 5-point Likert scale, with 1 representing "strongly disagree" and 5 representing "strongly agree." In responding to the first statement, relating to engagement, the students in the AI group agreed more strongly (Mean = 4.1, SD = 0.98) than those in the in-class active learning group (Mean = 3.6, SD = 0.92), $t(311) = -4.5$, $p < 0.0001$. Likewise, in responding to the second statement, relating to motivation, students in the AI group agreed more strongly (Mean = 3.4, SD = 1.0) than those in the in-class active learning group (Mean = 3.1, SD = 0.86), $t(311) = -3.4$, $p < 0.001$. Students' average level of agreement with the remaining two statements (relating enjoyment to growth mindset) 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 in-class active learning group, and the degree to which both groups enjoyed the lesson and reported a growth mindset was comparable.

Discussion

We have found that when students interact with our AI tutor, at home, on their own, they learn significantly more than when they engage with the same content during an in-class active learning lesson, 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 to the fullest extent current technology allows, thus establishing a model for future educational AI applications.

Designing successful student-AI interactions

A subset of the best practices (i-iii) were incorporated into the AI pedagogy by careful engineering of the AI tutor's system prompt. We designed the AI tutor with a system prompt with guidelines (Supplementary Material 1) 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), as the AI tutor would occasionally discuss parts out of sequence or that were not immediately relevant. For this reason, the AI platform was designed to guide students sequentially through each part of each problem in the lesson, mirroring the approach taken by the instructor during the in-class active learning (see screenshot of AI tutor platform in Figure S1).

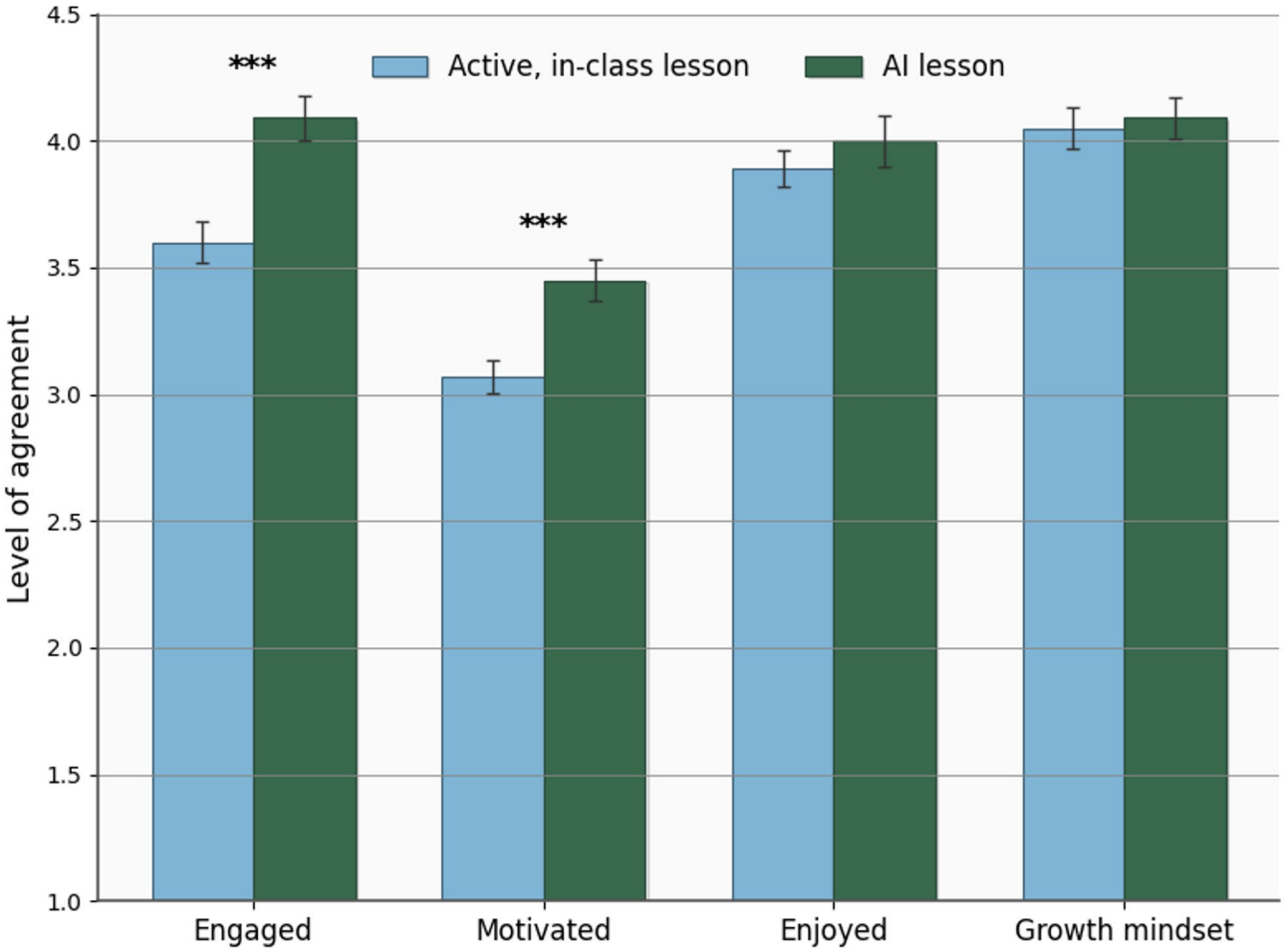

Fig. 3. Student perception of learning experiences: Level of agreement to statements about perceptions of learning experiences, comparing students taught by in-class active learning and students taught with the AI tutor. Error bars show 1 standard error of the mean. Asterisks above the bars denote *P*-values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

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 structured 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[15]. 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 in-class lesson. Students who are familiar with the material or have underlying skills, however, can move through the activities in less time than required for the in-class lesson. We measured the students' perception of pace during the control condition (in-class active learning) on the days the experiment took place. Notably, the 3.8% of students who found the pace of class "too fast" all spent more than the median time (49 minutes) on the AI lesson, while the 2.2% who found the pace of the in-class lesson "too slow" all spent less than the median time on the AI lesson.

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[2]. These previous approaches did not adhere to the same research-based best practices that informed our approach. Our success suggests that thoughtful implementation of AI-based tutoring could lead to significant

improvements to current pedagogy and enhanced learning gains for a broad range of subjects in a format that is accessible in 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 engagement with challenging material. AI could 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. As seen in previous studies[2,25], 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 in-class 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.

Context, limitations, and future directions

Our AI tutoring approach was applied in a setting where students were engaging substantially with material in particular subject areas for the first time. Our lessons were comprised of activities focused on learning objectives categorized at the understanding, applying, and analyzing levels of Bloom's Taxonomy[7]—as were the associated pre- and post-test questions[8]. This stage of learning, characterized by a meaningful degree of information delivery, appears to be particularly well suited for current generative AI tutors. The significant gains and positive affect observed in this study may also depend on several factors: a heterogeneous student population requiring varying instructional paces, integration of high-quality instructional videos,[9] a large language model capable of closely following complex prompts (e.g., GPT-4), expert-crafted, question-specific prompts written by instructors experienced with the content, a carefully structured framework designed to scaffold and guide student interactions, and content that lends itself to such a format. While the advantages of the experimental condition are widely generalizable and our findings have broad implications, we do not presume that structured AI tutoring will always outperform in-class active learning in all contexts, for example, those requiring complex synthesis of multiple concepts and higher-order critical thinking.

Compelling directions for future work include exploring other contexts throughout the learning process where AI tutoring may be successfully implemented, such as in homework, recitation, exam studying, pre-class assignments, and laboratory. Valuable follow-up studies could also explicitly examine the details of such combinations throughout an entire course. This would also allow for systematic integration of well-established retention enhancing strategies (e.g., spacing) and could provide insights into other novel phenomena that may arise from prolonged and varied use of AI in education, such as potential impact on collaboration skills. Given that the current AI tutor implementation mirrors well-established in-class active learning pedagogies and generates comparable affect[10]—with its primary difference (besides personalization) being the medium of delivery, which typically does not impact learning on its own[26]—it is reasonable to expect findings from in-class active learning approaches to carry over. Nonetheless, studies that explicitly replicate known in-class active learning results[27,28] would be valuable for confirming and refining the details of this transferability. Such research could include explorations of the qualities that constitute effective system prompts and behaviors for AI tutors in various situations (e.g., determining when the AI tutor should openly provide answers versus guiding students to reflect on their own responses).

Generative AI technology is developing very rapidly, allowing for expansion of the capabilities and application of AI tutoring. While accuracy of our AI tutor relied on pre-written answers, as generative AI models improve in scientific reasoning[11], studies could explore whether such efficacy could be achieved without a provided solution.

[7] Bloom's Taxonomy is a hierarchical model used to classify educational learning objectives into levels of complexity and specificity. The original taxonomy was revised in 2001 and is as follows: remembering, understanding, applying, analyzing, evaluating, and creating. Anderson, L. W. & Krathwohl, D. R. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Longman, 2001).

[8] 33% analyzing, 41% applying, 21% understanding, and 4% remembering.

[9] Videos were produced at the Harvard University Derek Bok Center production studio, and the instructor (GK) has a decade of experience hosting, writing, and producing videos and documentaries (e.g., via NOVA | PBS).

[10] Equivalent growth mindset and enjoyment, increased engagement and motivation, and improved satisfaction with feedback.

[11] Currently, the models with the most advanced scientific reasoning capabilities have longer response times; in the context of AI tutoring, the model choice should take into account efficiency as well as reasoning capabilities.

Also, in our approach, feedback was provided in response to student input, but multimodality would allow AI systems to interpret images (or audio) of a student's work and more proactively provide feedback. Investigations could explore whether holistic monitoring of a student's process could provide feedback on issues with thinking that may not be addressed by current pedagogies (in or out of the classroom) in which students typically receive targeted feedback only when they ask a question.

Conclusion

We have built an AI-based tutor, engineered with appropriate prompts and scaffolding, that helps students learn significantly more in less time and feel more engaged and motivated compared with in-class active learning. 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 frameworks that are engaging and suggest a pathway for widely accessible education on which policymakers, technologists, and educators can collaborate. It also serves as a foundation for a series of explorations of the use of AI in educational settings.

Methods

Study population

The present study took place during 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 S2A), as were previous measures of their physics background knowledge (see Table S2B). Note that FCI pretest scores are comparable to those of students at other universities[29]. 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.

We note that the results hold for students with both lower and higher performance abilities than typical students at various institutions, as assessed by the widely recognized FCI test. They also hold for all degrees of scientific attitudes—from non-expert to expert—as gauged by the CLASS survey. Typical pre-instruction FCI scores in educational institutions range from aproximately 30% to 50%[30]. In this study, significant subpopulations span this range. The subpopulations of students with pre-instruction FCI scores below 40% and those above 40% both showed significantly better post-test performance with AI tutoring compared to in-class active learning ($p < 0.001$). This improvement was similarly observed in students below and above the 65% benchmark[31] in scientific attitudes on the CLASS.

Course setting

The course (PS2) meets twice per week for 75 minutes each. The study took place during one of the two meeting of the class during the ninth and tenth weeks of the course. All in-class lessons employed research-based best practices for in-class active learning[32]. 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 difficulties.

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 [33]. We note that the authors and instructors represented in the literature demonstrating this in-class active learning approach to be effective and superior to passive instruction overlap with those of the present study. Similar active learning approaches have been shown to increase learning across a wide range of STEM fields[34]. Although active learning pedagogies may elicit negative perceptions from students[35], both course instructors, as well as their presentations in the course, achieved student evaluation scores above the departmental and division averages. The active learning pedagogy, optimized over the years to cater to the vast majority of the student population, ensures minimal inactivity by strategically allowing instructors to intervene and adjust the class dynamically based on real-time observations of student engagement and performance[32].

To verify the active learning emphasis of the class, students were asked the following question 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 crossover design. The design allowed for control of all aspects of the lessons that were not of interest. The crossover design is summarized in Table S3. 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. 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 crossover 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 both conditions 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 in order to receive participation credit they needed to demonstrate that they had made an honest effort in completing the tests.

Additional controls

In addition to using a crossover 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 versus 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.

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 engaging in test conversations for each question and iterating. The most significant time commitment involved in preparing the AI-supported lessons was the development of an AI tutor platform software that took pedagogical best practices into consideration (e.g., structured around individual questions embedded in individual assignments), which took several months.

Data availability

https://github.com/HarvardAItutor/Study-Data-v4.

Received: 25 March 2025; Accepted: 7 April 2025
Published online: 03 June 2025

References

1. Singer, N. Will chatbots teach your children? *New York Times* https://www.nytimes.com/2024/01/11/technology/ai-chatbots-khan-education-tutoring.html (2024).
2. Forero, M. G. & Herrera-Suárez, H. J. ChatGPT in the classroom: boon or bane for physics students' academic performance? Preprint at arXiv:2312.02422 (2023).
3. Kumar, H., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. Math education with large language models: peril or promise? *SSRN* https://doi.org/10.2139/ssrn.4641653 (2023).
4. Sweller, J. Cognitive load theory. In *The Psychology of Learning and Motivation: Cognition in Education* (eds Mestre, J. P. & Ross, B. H.) 37–76 (Elsevier Academic Press, 2011).
5. Kortemeyer, G. Could an artificial-intelligence agent pass an introductory physics course?. *Phys. Rev. Phys. Educ. Res.* **19**(1), 010132 (2023).
6. Henderson, C. & Dancy, M. H. Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Phys. Rev. Spec. Top.-Phys. Educ. Res.* **3**(2), 020102 (2007).
7. Stains, M. et al. Anatomy of STEM teaching in North American universities. *Science* **359**(6383), 1468–1470 (2018).
8. Handelsman, J. et al. Scientific teaching. *Science* **304**, 521–522 (2004).
9. Hake, R. R. 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. Crouch, C. H. & Mazur, E. Peer instruction: Ten years of experience and results. *Am. J. Phys.* **69**, 970–977 (2001).
11. Deslauriers, L., Schelew, E., & Wieman, C. Improved learning in a large-enrollment physics class. *Science* **332**, 862–864 (2011).
12. Freeman, S. et al. Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8410–8415 (2014).
13. Fraser, J. M. et al. Teaching and physics education research: Bridging the gap. *Rep. Prog. Phys.* **77**, 032401 (2014).

14. Thornton, R. K., Kuhl, D., Cummings, K. & Marx, J. Comparing the force and motion conceptual evaluation and the force concept inventory. *Phys. Rev. ST Phys. Educ. Res.* **5**, 010105 (2009).
15. Bloom, B. S. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res.* **13**(6), 4–16 (1984).
16. Dweck, C. S. *Mindset: The New Psychology of Success* (Random House, 2006).
17. Yeager, D. S. & Dweck, C. S. What can be learned from growth mindset controversies?. *Am. Psychol.* **75**(9), 1269 (2020).
18. Hestenes, D., Wells, M. & Swackhamer, G. Force concept inventory. *Phys. Teach.* **30**(3), 141–158 (1992).
19. Fredricks, J. A., Blumenfeld, P. C. & Paris, A. H. School engagement: Potential of the concept, state of the evidence. *Rev. Educ. Res.* **74**(1), 59–109. https://doi.org/10.3102/00346543074001059 (2004).
20. Wood, D., Bruner, J. S., & Ross, G. The role of tutoring in problem-solving. *J. Child Psychol. Psychiatry* **17**(2), 89–100 (1976).
21. Shute, V. J. Focus on formative feedback. *Rev. Educ. Res.* **78**(1), 153–189 (2008).
22. Tatum, B. C. & Lenel, J. C. A Comparison of self-paced and lecture/discussion methods in an accelerated learning format. *J. Res. Innov. Teach.* **5**(1), (2012).
23. Meyer, J. G. et al. ChatGPT and large language models in academia: Opportunities and challenges. *BioData Min.* **16**(1), 20–20. https://doi.org/10.1186/s13040-023-00339-9 (2023).
24. Nye, M. et al. Show your work: Scratchpads for intermediate computation with language models. arXiv:2112.00114 (2021).
25. Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J., Unreflected acceptance—investigating the negative consequences of ChatGPT-assisted problem solving in physics education. In *HHAI 2024: Hybrid Human AI Systems for the Social Good, Frontiers in Artificial Intelligence and Applications*, Vol. 386, 199–212. IOS Press (2024).
26. Clark, R. C. & Mayer, R. E. e-Learning and the science of instruction. In *E-learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*. Center for Creative Leadership (2011).
27. Francis, G. E., Adams, J. P. & Noonan, E. J. Do they stay fixed?. *Phys. Teach.* **36**, 488–490 (1998).
28. McDermott, L. C., Heron, P. R. L., Shaffer, P. S., & Stetzer, M. S. Improving the preparation of K-12 teachers through physics education research. *Am. J. Phys.* **74**, 763–767 (2006).
29. Caballero, M. D. et al. Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study. *Am. J. Phys.* **80**(7), 638–644 (2012).
30. Hake, R. R. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.* **66**, 64–74 (1998).
31. Perkins, K. K., Adams, W. K., Pollock, S. J., Finkelstein, N. D., & Wieman, C. E. Correlating student beliefs with student learning using the Colorado Learning Attitudes about Science Survey. *AIP Conf. Proc.* **790**, 61–64 (2005).
32. McCarty, L. S. & Deslauriers, L. Transforming a large university physics course to student-centered learning, without sacrificing content: A case study. In *The Routledge International Handbook of Student-Centered Learning and Teaching in Higher Education*, 186–200 (Routledge, 2020).
33. Miller, K., Callaghan, K., McCarty, L. S. & Deslauriers, L. Increasing the effectiveness of active learning using deliberate practice: A homework transformation. *Phys. Rev. Phys. Educ. Res.* **17**(1), 010129 (2021).
34. Freeman, S. et al. Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci.* **111**(23), 8410–8415 (2014).
35. Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K. & Kestin, G. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proc. Natl. Acad. Sci.* **116**(39), 19251–19257 (2019).

Acknowledgements

Acknowledgments: We wish to thank Logan McCarty for thoughtful comments, conversations, insights, and edits as well as for his 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 University 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: G.K., K.M., A.K., T.W.M. Methodology: G.K., K.M., A.K., G.P. Software Conceptualization and Design: G.K. Software Engineering: G.K. Project Administration: G.K. Validation: G.K., K.M. Formal Analysis: G.K., K.M. Investigation: G.K., K.M., T.W.M., G.P. Data Curation: G.K., K.M. Writing—Original Draft: G.K., K.M. Writing—Review & Editing: G.K., K.M., A.K., G.P. Supervision: G.K., K.M.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-97652-6.

Correspondence and requests for materials should be addressed to G.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025