Harvard AI tutor against active learning in physics

IMPLEMENTED 2023–2025 · EDUCATION · US · BAND D · PARTLY PROVED
VERSION 1.0 · PUBLISHED 15 NOVEMBER 2026 · UPDATED 15 NOVEMBER 2026 · FACTS CHECKED 15 NOVEMBER 2026

Harvard's physics teaching team built an AI tutor for an introductory physics course and tested it against an in-class active learning lesson in a randomised crossover trial in autumn 2023. The peer-reviewed paper found larger learning gains with the tutor, in less time. The test covered two lessons in one course, and the authors do not claim the result holds for every context.

ABOUT THIS SYNOPSIS

This is MarLuc's summary and grading of publicly available sources, for information only. It is not investment, legal, financial or professional advice, and should not be relied on for any decision. Readers must validate it against the original sources and, where appropriate, challenge its conclusions.

Facts and figures are as reported by the sources listed below, which may be incomplete, self-reported or out of date. Check the original source before relying on any figure.

The impact band, proof marker and SEAM view are MarLuc's opinion. The organisation is not a MarLuc client and has not reviewed or endorsed this synopsis. Names and trademarks belong to their owners.

NOTICE WORDING LAST UPDATED 16 NOVEMBER 2026

AT A GLANCE

AI used: Generative AI tutor: a custom chatbot built on OpenAI's GPT-4, with instructor-written prompts and step-by-step solutions.

Implemented: 2023–2025. Grade: band D, Task support, Partly proved.

Headline result, as reported: In a randomised crossover trial of 194 Harvard students, a purpose-built GPT-4 tutor produced over double the median learning gains of an in-class active learning lesson, in a median of 49 minutes against 60

1 · The case

The facts, as reported by the 2 sources listed at the end. Where a source is silent, the field says "Not reported".

Organisation

Harvard University, Physical Sciences 2 teaching team (Kestin, Miller and colleagues). Harvard University, Physical Sciences 2, an introductory physics course for life sciences majors and, per the paper, Harvard's largest physics class (233 students enrolled). Canvas block: 5. People and skills. Also 4. Operations and process.

Business challenge

The challenge. Per the authors, passive lectures remain in wide use in STEM courses, and even active learning, with one teacher working with many students, struggles to give personal feedback and keep pace right for everyone. Earlier studies of AI chatbots in teaching had shown mixed results when students used them without guidance.

How it was addressed. The team built a custom tutor on GPT-4 with prompts written by instructors, including step-by-step solutions for each activity. Students were randomly assigned to two groups and each experienced two lessons in consecutive weeks, one in class with active learning and one at home with the AI tutor, with the conditions reversed in the second week. Content and worksheets were identical. [1] [2]

Who implemented it

In-house (In-house): Designed the tutor, wrote the prompts and ran the trial in its own course..
OpenAI (Model or cloud provider): Supplied the GPT-4 model that the tutor is built on..
The Gazette says the Derek Bok Center and Harvard University IT are working on wider pilots.

Technology (SEAM pillar E)

SEAM types: T3 API integrations. Deployment: Cloud APIs. Vendors and products: OpenAI, GPT-4 via the OpenAI API.

Where it was built in: A website built by the teaching team that hosts the tutor. Each lesson used a system prompt plus content-rich prompts with step-by-step solutions written by instructors; students used it at home on their own.

What changed

Before: new physics material was introduced in a 75-minute class session using active learning.
In the trial: for one lesson each, students met the same material through the AI tutor at home.
The course's regular teaching was not replaced. [1] [2]

What it cost

Not reported

What it saved

Time: median 49 minutes on the AI lesson against 60 minutes assumed for the in-class lesson.
Learning: median post-test 4.5 against 3.5, with median learning gains over double. Engagement and motivation were higher with the tutor; enjoyment and growth mindset were comparable.
Cost saving: Not reported. [1] [2]

Governance and control

IRB approval, a randomised crossover design with an in-class control, and peer review. The authors say the tutor should not replace in-class instruction and warn that AI can undermine learning if students use it as a crutch.

S2 · Accountable owner · Claimed

Lecturer Gregory Kestin and senior lecturer Kelly Miller led the study, per the Gazette and the paper's author list. Who owns the tutor after the trial: Not reported. [1] [2]

A1 · Safe use · Evidenced
The paper says the team did not rely on GPT-4 alone for solutions and added step-by-step answers to the prompts to limit hallucinations; 83% of students rated the tutor's explanations as good as or better than the class instructors'. [1]

A2 · Human in the loop · Claimed
The Gazette says conversations, including the tutor's personality and feedback quality, are pre-vetted, and the authors advise against AI replacing in-class instruction. [1] [2]

A3 · Monitoring and assurance · Evidenced
Students took pre-tests and post-tests for each lesson; the team tracked time on task and ran regression and quantile analyses. [1]

A4 · Vendor AI · Not reported
Contract terms, data handling and vendor controls for the OpenAI API are not reported in the sources opened.

A5 · Agent controls · Not reported
No agents described.

A6 · Standards and regulation · Evidenced
The study was approved by the Harvard University IRB (study no. IRB23-0797) and passed peer review at Scientific Reports. [1]

What worked, what did not, what is unproven

WORKED

– Median post-test score 4.5 with the tutor against 3.5 in class, a significant difference (z = -5.6, p < 10^-8).

– The authors give an effect size of 0.63 from a linear model and 0.73 to 1.3 standard deviations from a quantile regression that corrects for the ceiling effect.

– Students spent a median of 49 minutes on the AI lesson against 60 minutes assumed for the class lesson.

– Engagement (4.1 against 3.6) and motivation (3.4 against 3.1) were higher with the tutor.

DID NOT

– Enjoyment and growth mindset showed no significant difference between the two conditions.

– The authors do not presume that structured AI tutoring will always outperform in-class active learning, for example where complex synthesis and higher-order critical thinking are needed.

– The authors say results may depend on factors including a heterogeneous student population, high-quality videos, a capable model such as GPT-4 and expert-written prompts.

UNPROVEN

– Whether the gains last: the paper reports post-tests after each lesson and suggests retention strategies for future work.

– Results over a full course, in other subjects or at other institutions: the authors call for follow-up studies.

– Whether the tutor is in routine use at Harvard: Not reported; the Gazette (September 2024) describes pilots in other courses.

– Dollar cost of building and running the tutor: Not reported.

Timeline

2023 Trial runs in the Fall 2023 semester in Physical Sciences 2, with an IRB-approved crossover

Sept 2024 Harvard Gazette reports preliminary findings; the Derek Bok Center and Harvard University IT plan to pilot similar chatbots in a handful of large introductory courses that autumn [2]

Apr 2025 Scientific Reports accepts the paper [1]

Jun 2025 Paper published open access in Scientific Reports [1]

2 · Grade and SEAM assessment

MarLuc's opinion, made on the evidence above. How cases are graded: marluc.co.uk/case-studies/how-we-rate

Level of change and impact

Claimed: . Grade: D, Task support. Proof: Partly proved. Level of change: Task support.

Task support (Level 1, band D). The trial tested one tutor in two lessons. The sources opened do not show any course or process redesigned around it.

The trial shows a tutor doing one job well: introducing new material to students outside class. Harvard's own account says other courses are piloting similar tutors, but the sources opened do not show any course redesigned around it, so this is task support (Level 1, band D). The evidence is strong for what was tested and limited in scope.

The SEAM view and lessons learnt

This is a rare randomised comparison of an AI tool against a strong human method, in a live course. The tutor won where it was carefully designed: instructor-written prompts, step-by-step solutions and a lesson matched to the class version. The paper says unguided chatbots did worse in earlier studies, so the design work is part of the result. For a larger organisation, the transferable point is to test an AI tool against the current best method, in the real setting, before deciding to scale it.

1. Compare the AI tool with your best current method, not with nothing.
2. Put expert-written content and checks into the prompts before students use it.
3. Keep the human class time and use the tool where it helps most, such as first contact with new material.

Evidence

Score 3 of 4. Independent: peer-reviewed randomised controlled trial (Scientific Reports, 2025). Limits: authors are also the course's teachers and tutor designers, two lessons, one course at one university, no long-term follow-up.

Evaluation method. Randomised crossover trial in a live course: students randomly assigned to two groups, each experiencing an in-class active learning lesson and an AI tutor lesson in consecutive weeks, with pre-tests and post-tests and a four-question perception survey.

Sources disagree. The Gazette describes 'double' learning gains and 'engagement doubled' in its headline; the paper reports median learning gains 'over double' and engagement ratings of 4.1

3 · Original sources

Read these in full. To challenge a fact or a grade, use the contact form at marluc.co.uk or the link on https://marluc.co.uk/case-studies/harvard-ai-tutor-physics-trial

[1] Scientific Reports: AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting
https://www.nature.com/articles/s41598-025-97652-6

[2] Harvard Gazette: Professor tailored AI tutor to physics course. Engagement doubled.
https://news.harvard.edu/gazette/story/2024/09/professor-tailored-ai-tutor-to-physics-course-engagement-doubled/

Harvard University, Physical Sciences 2 teaching team (Kestin, Miller and colleagues) website: https://www.harvard.edu

Version 1.0 · Published 15 November 2026 · Updated 15 November 2026 · Facts checked against sources 15 November 2026 · Downloaded 26 November 2026. This PDF may be out of date: the latest version, and a way to report an error or a better source, is at https://marluc.co.uk/case-studies/harvard-ai-tutor-physics-trial

VERSION HISTORY

v1.0 15 November 2026 First published