Unguarded GPT-4 in Turkish maths classes (Bastani et al.)

IMPLEMENTED 2023–2024 · EDUCATION · TURKEY · BAND F · PARTLY PROVED
VERSION 1.0 · PUBLISHED 15 NOVEMBER 2026 · UPDATED 15 NOVEMBER 2026 · FACTS CHECKED 15 NOVEMBER 2026

Researchers gave nearly 1,000 high school students in Turkey GPT-4 tutors during maths practice sessions. Students with an unguarded chat tool did better in practice and then scored lower on the unaided exam than students with no AI. A tool built to give hints, with teacher-supplied solutions, avoided the drop but did not improve exam scores.

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Facts and figures are as reported by the sources listed below, which may be incomplete, self-reported or out of date.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

What went wrong: Weak oversight or controls.

AI used: Generative AI tutor: two GPT-4 chat tools built by the researchers, one like ChatGPT (GPT Base) and one with teacher-designed hints and solutions in the prompt (GPT Tutor).

Implemented: 2023–2024. Grade: band F, Failure, Partly proved.

Headline result, as reported: In a preregistered randomised trial at one Turkish high school with nearly 1,000 students, unrestricted GPT-4 access raised practice scores by 48% but lowered later unaided exam scores by 17%; a guarded tutor removed the harm but showed no gain

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

University of Pennsylvania researchers with a Turkish high school. Field experiment by researchers at the University of Pennsylvania's Wharton and engineering schools in a large Turkish high school, covering about fifty grade 9 to 11 classes and nearly 1,000 students. Canvas block: 8. Governance and risk. Also 3. Customers and proposition.

Business challenge

The challenge. Generative AI raises performance on tasks, but the authors ask whether it harms how people learn skills, given that it can be wrong and users need the skill to check it.

How it was addressed. Classes were randomly assigned to one of three arms during review sessions. One had textbooks and notes only. One had a GPT-4 chat tool like ChatGPT. One had a GPT-4 tool told to give hints without answers and loaded with teacher-written solutions. After practice, all students sat a closed-book exam on conceptually similar problems. [1] [2]

Who implemented it

University of Pennsylvania (University or research lab): Designed and ran the field experiment and built the tutoring tools.
OpenAI (Model or cloud provider): GPT-4 powered both tutor variants.
The high school is described only as a large high school in Turkey.

Technology (SEAM pillar E)

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

Where it was built in: The researchers built a custom maths tutoring chat interface on GPT-4. Students used it on provided laptops in class during the practice phase of each session.

What changed

No lasting change at the school is reported. In the experiment, students with the unguarded tool did better in practice and worse in the unaided exam; the guarded tool removed the exam penalty without a gain. [1] [2]

Cost of failure

Not reported in money. Measured harm: unaided exam scores 17% below the control arm for students who had the unguarded tool. Both practice and exam scores counted towards students' final grades, per the paper.

What it saved

Not reported. [1] [2]

Governance and control

IRB approval, preregistration, teacher-informed guardrails, classroom supervision and independent grading, per the paper. Vendor terms and data handling: Not reported.

S2 · Accountable owner · Evidenced
The study was approved by the University of Pennsylvania IRB and deemed exempt; the authors name the study team in the paper. [1]

A1 · Safe use · Evidenced
The guarded variant was built with teacher input to hold back answers and supply correct solutions. [1]

A2 · Human in the loop · Evidenced
A teacher and a staff member were present in each experimental class; exams were unaided and independently graded. [1]

A3 · Monitoring and assurance · Evidenced
The trial was preregistered with the unaided exam as the primary analysis, and independent graders marked papers. [1]

A4 · Vendor AI · Not reported
Not reported.

A5 · Agent controls · Not reported
Not reported.

A6 · Standards and regulation · Not reported
Not reported.

What worked, what did not, what is unproven

WORKED

– Practice scores rose 48% with GPT Base and 127% with GPT Tutor against control.

– The GPT Tutor safeguards largely removed the exam penalty (-0.004, not distinguishable from control).

– Students used GPT Tutor in more substantive ways, such as asking for help or trying answers, than they used GPT Base.

DID NOT

– Students with GPT Base scored 0.054 lower than control on the unaided exam, a 17% fall, per the paper.

– GPT Tutor did not raise exam scores above control.

– Students did not perceive any loss of learning from copying solutions, per the authors.

UNPROVEN

– Long-term effects, which the authors say were out of scope because of limits set by the partner school.

– Results for other subjects, tutor designs, models and countries; the authors note the models have improved since Fall 2023.

– The content of the August 2025 PNAS correction, which the sources opened do not describe.

Timeline

2023 Fall semester field experiment begins at a large Turkish high school [1]

Nov 2024 Paper received by PNAS [1]

May 2025 Paper accepted [1]

Jul 2025 Published in PNAS volume 122, issue 26 [2]

Aug 2025 PNAS publishes a correction notice to the article; its content was not opened [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: Not applicable. Grade: F, Failure. Proof: Partly proved. Level of change: Task support.

No level assessed: band F, measured harm to later unaided performance in an experimental

The SEAM view and lessons learnt

The result separates two things a buyer might treat as one: performance with the tool and capability without it. The same model produced opposite effects depending on prompt design and teacher input. For a larger organisation, the transferable point is to test what staff can do when the tool is absent, and to design guardrails before wide rollout. It is one school and four sessions, so it is a warning about design rather than a verdict on AI tutoring.

1. Measure unaided performance as well as assisted performance.
2. Build guardrails and checked answers into the tool before rollout.
3. Test each design change against a comparison group.

Evidence

Score 3 of 4. Independent: peer-reviewed, preregistered randomised trial in PNAS with independent graders. Limits: one school, one subject, short term.

Evaluation method. Randomised controlled trial at classroom level with three arms, preregistered primary analysis of unaided exam scores, intention-to-treat regression.

Sources disagree. None found between the two sources opened. Secondary press coverage, not opened, was seen to describe the 17% figure against different comparison groups; the paper states it relative to the control arm.

Status. Published in PNAS July 2025; a correction notice followed in August 2025.

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/bastani-turkish-maths-unguarded-gpt4

[1] PNAS via Europe PMC: Generative AI without guardrails can harm learning: Evidence from high school mathematics (Bastani et al., 2025)
https://www.ebi.ac.uk/europepmc/webservices/rest/PMC12232635/fullTextXML

[2] IDEAS/RePEc: Generative AI without guardrails can harm learning: Evidence from high school mathematics (PNAS 122(26))
https://ideas.repec.org/a/nas/journl/v122y2025pe2422633122.html

University of Pennsylvania researchers with a Turkish high school website: https://www.pnas.org

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/bastani-turkish-maths-unguarded-gpt4

VERSION HISTORY