Stanford Tutor CoPilot randomised trial

IMPLEMENTED 2024 · EDUCATION · US · BAND D · PARTLY PROVED
VERSION 1.0 · PUBLISHED 15 NOVEMBER 2026 · UPDATED 15 NOVEMBER 2026 · FACTS CHECKED 15 NOVEMBER 2026

Stanford researchers built an AI assistant that suggests expert-style responses to human tutors during live maths tutoring and tested it in a randomised trial. Students whose tutors had access passed more lesson exit tickets, with the largest gains for lower-rated tutors. Students' end-of-year test scores showed no significant difference, and the study is a working paper.

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Facts and figures are as reported by the sources listed below, which may be incomplete, self-reported or out of date.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 assistant for tutors: Tutor CoPilot, built on GPT-4, suggests hints, questions and explanations to a human tutor during live chat tutoring.

Implemented: 2024. Grade: band D, Task support, Partly proved.

Headline result, as reported: In a preregistered randomised trial with FEV Tutor and a US school district, students whose tutors could use an AI assistant were 4 percentage points more likely to pass lesson exit tickets (62% to 66%); the authors report no significant effect on end-of-year tests

1 · The case

The facts, as reported by the 3 sources listed at the end. Where a source is silent, the field says "Not reported".

Organisation

Stanford University, with FEV Tutor and a US school district. Stanford researchers working with FEV Tutor, a virtual tutoring provider, in a large southern US school district serving more than 30,000 students. Nine schools took part; students were in grades 3 to 8 and below grade level on the state test. Canvas block: 4. Operations and process. Also 5. People and skills.

Business challenge

The challenge. Per the authors, novice tutors often give generic feedback or give away the answer, expert training is costly and hard to scale, and students in under-served communities are most likely to be taught by inexperienced educators.

How it was addressed. Tutor CoPilot is a button inside the tutoring platform. A tutor can press it during a session to receive a suggested response, such as a hint or a guiding question, generated by a language model prompted with expert reasoning. Tutors were randomly assigned to have the tool or not, and treatment tutors were trained to use it. [1] [2] [3]

Who implemented it

Stanford University (University or research lab): Built Tutor CoPilot and ran the randomised trial.
FEV Tutor (Industry partner): Virtual tutoring provider whose platform hosted the tool and whose tutors took part.
OpenAI (Model or cloud provider): GPT-4 generated the suggestions, per The 74.
The paper names the school district only as a large southern US district; The 74 says the researchers prompted GPT-4 to generate the guidance.

Technology (SEAM pillar E)

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

Where it was built in: A Tutor CoPilot button was added to a virtual tutoring platform with a problem display, whiteboard and chat. When pressed, it sends the de-identified recent conversation and the lesson topic to a language model and shows the tutor a suggested response that the tutor can edit.

What changed

Before: tutors responded unaided. In the trial: treatment tutors had on-demand suggestions. Students of treatment tutors passed more exit tickets and tutors used more guiding strategies, per the authors' classifiers. No lasting operational change is reported. [1] [2] [3]

What it cost

About $20 per tutor a year, by the authors' estimate from tutors' usage during the study.

What it saved

Not a saving. The authors compare $20 a year with professional development costing $3,300 per teacher or more annually. [1] [2] [3]

Governance and control

De-identified student and tutor names, a limit of 10 messages sent to the external service, and human choice over each suggestion, per the paper. Accountable owner and vendor terms: Not reported.

S2 · Accountable owner · Claimed
The lead author worked part-time with FEV Tutor and collaborated with its engineering, tutoring and curriculum teams; no single accountable owner is named. [1]

A1 · Safe use · Claimed
The paper says the system de-identifies names and limits the context sent to external services; the authors note identity can still be inadvertently disclosed. [1]

A2 · Human in the loop · Evidenced
The tutor decides what to say. Suggestions can be edited, regenerated or replaced, and tutors told the researchers some were not grade-level appropriate. [1]

A3 · Monitoring and assurance · Evidenced
Randomisation, a preregistered analysis plan, tutor interviews and balance and attrition checks are documented in the paper. [1]

A4 · Vendor AI · Not reported
Vendor terms for the external language model service are not reported. [1]

A5 · Agent controls · Not reported
Not reported.

A6 · Standards and regulation · Not reported
Not reported.

What worked, what did not, what is unproven

WORKED

– Students of treatment tutors were 4 percentage points more likely to pass exit tickets (62% to 66%, p<0.01).

– Students of lower-rated tutors passed 65% of exit tickets against 56% in control, a 9 point gain.

– Treatment tutors were more likely to ask guiding questions and less likely to give away the answer, per the authors' classifiers.

DID NOT

– The authors found no statistically significant improvement in end-of-year maths test scores.

– About 29% of treatment sessions used the tool at all.

– Tutors said suggestions were sometimes too advanced for the grade level and needed simplifying.

UNPROVEN

– Effects on lasting learning; the authors say exposure varied little and the study was short.

– Whether the $20 cost holds once training is counted. Benjamin Riley, quoted by The 74, said tutors received up to three weeks of training.

– Results in other districts, subjects, countries or with more affluent students.

Timeline

Mar 2024 Study starts in nine schools of a large southern US school district and runs for two months [1]

Oct 2024 First version of the paper posted on arXiv [1]

Jan 2025 Version 2 posted with 900 tutors assigned and 1,800 students described [1]

2025 National Student Support Accelerator lists the study with a revised abstract of more than 700 tutors and 1,000 students [2]

Feb 2026 The NSSA page links FutureEd coverage of two newer randomised trials of AI in live tutoring [2]

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

Task support (Level 1, band D). The AI supports individual tutors in the moment; the tutoring process and roles are unchanged.

Graded D: the AI assists individual tutors with in-session responses; the gain is modest and measured on a short-term proxy, and nothing in the sources shows a changed tutoring process, role or budget.

The SEAM view and lessons learnt

This is a human-plus-AI design: the person stays accountable and the tool lifts the weakest performers most. The effect is modest and measured on exit tickets, with no significant end-of-year change in the sources opened. For a larger organisation, the transferable idea is to put suggestions inside the existing workflow and measure who gains most, while counting training in the cost.

1. Put the AI inside the existing workflow and keep the person in charge.
2. Measure who gains most as well as the average.
3. Count training and support in the cost, not just the licence.

Evidence

Score 2 of 4. Own evaluation with preregistered randomisation: the tool's developers ran the trial, reported in a preprint and working paper. Not independently replicated; no journal publication found.

Evaluation method. Preregistered randomised controlled trial at tutor level, intention-to-treat analysis of session-level exit tickets, classifier analysis of 550,000+ messages and tutor interviews.

Sources disagree. The arXiv paper (v2) describes 900 tutors and 1,800 students; the National Student Support Accelerator abstract describes more than 700 tutors and 1,000 students and more than 350,000 messages. The paper itself gives 874 tutors in the schools, 1,787 students identified, and 782 tutors at launch. The 74 describes about 1,000 students and 900 tutors.

Status. Study complete (2024). Further rollout by FEV Tutor is not reported in the sources opened.

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/stanford-tutor-copilot-trial

[1] arXiv: Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise (Wang et al., v2, January 2025)
https://arxiv.org/pdf/2410.03017v2

[2] National Student Support Accelerator, Stanford: Tutor CoPilot study page
https://nssa.stanford.edu/studies/tutor-copilot-human-ai-approach-scaling-real-time-expertise

[3] The 74: Study: AI-Assisted Tutoring Boosts Students' Math Skills
https://www.the74million.org/article/study-ai-assisted-tutoring-boosts-students-math-skills/

Stanford University, with FEV Tutor and a US school district website: https://nssa.stanford.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/stanford-tutor-copilot-trial

VERSION HISTORY

v1.0 15 November 2026 First published