The Effect of Tutor CoPilot for Virtual Tutoring Sessions

Testing an AI Intervention to Improve Tutor Instruction

Rose E. Wang
Stanford Computer Science
PhD Candidate

Website: https://cs.stanford.edu/~rewang
Twitter: @rose_e_wang

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Scaling high-dosage tutoring is key to closing achievement gaps

The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence

Andre Nickow[1], Philip Oreopoulos[2], and Vincent Quan[3]

"We find that tutoring yields consistently substantial ***positive impacts on learning****"*

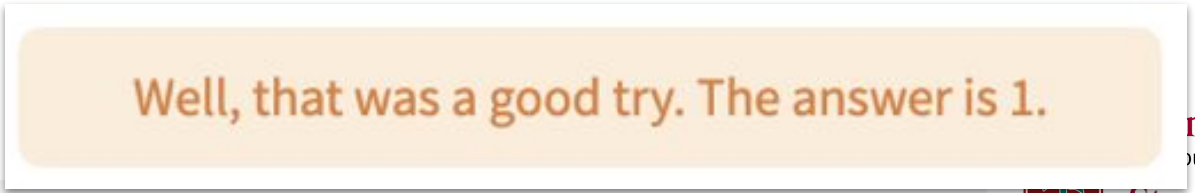
Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Remediating mistakes is key to student learning, but novice tutors struggle with remediation.

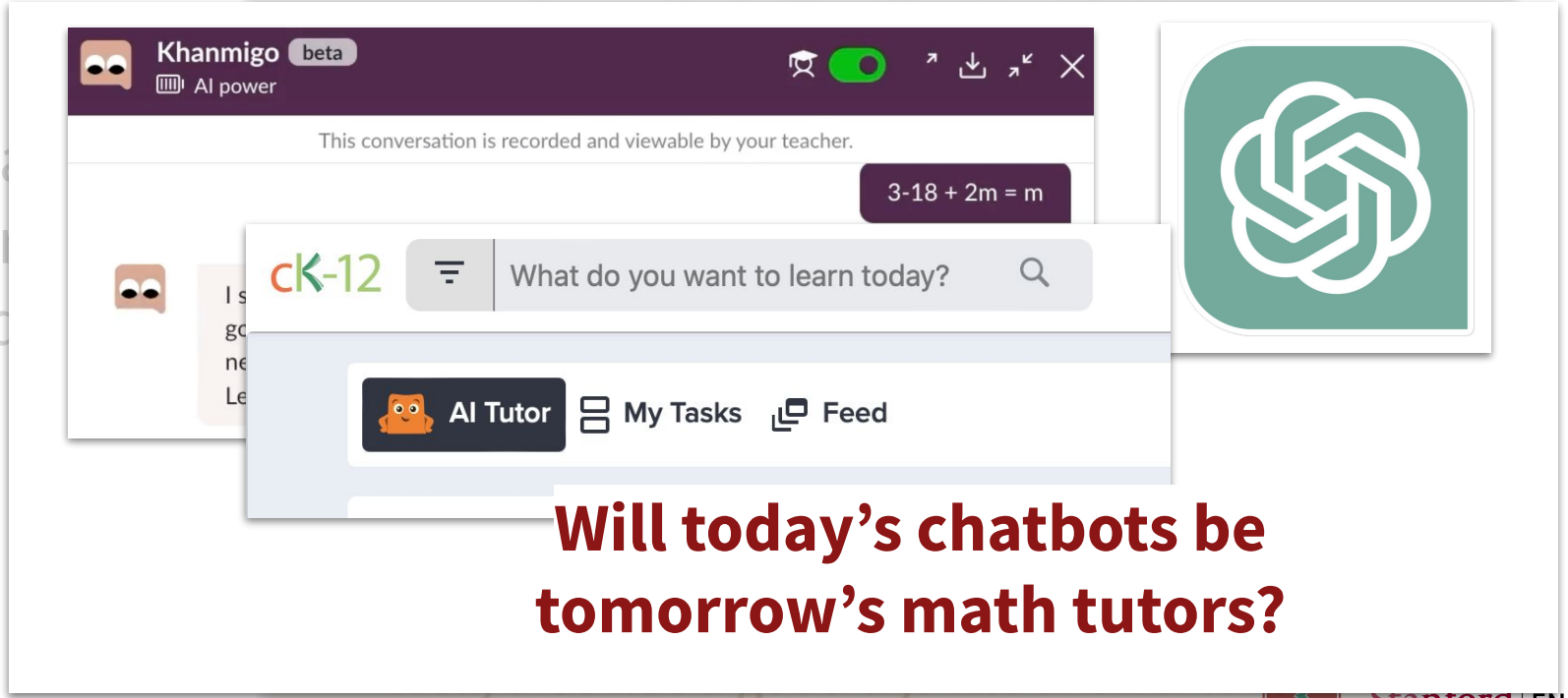
Imagine you're back in 3rd grade!

Novice tutors **don't engage** with the student.



Rose E. Wang (rewang@cs.stanford.edu)

Will tutoring quality be improved with auto-tutors, like with ChatGPT and large language models (LLMs)?



Will today's chatbots be tomorrow's math tutors?

Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Out-of-the-box LLMs do not improve pedagogical quality.

Large language models **do not engage** with the student out-of-the-box.

Actually, if Mike had 4 cookies and ate 3, we need to subtract. The answer is 1.

References:

Is ChatGPT a Good Teacher Coach? Rose E. Wang and Dorottya Demszky. BEA 2023. 🏆

Bridging the Novice-Expert Gap via Models of Decision Making. Rose E. Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, Dorottya Demszky. NAACL 2024.

Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

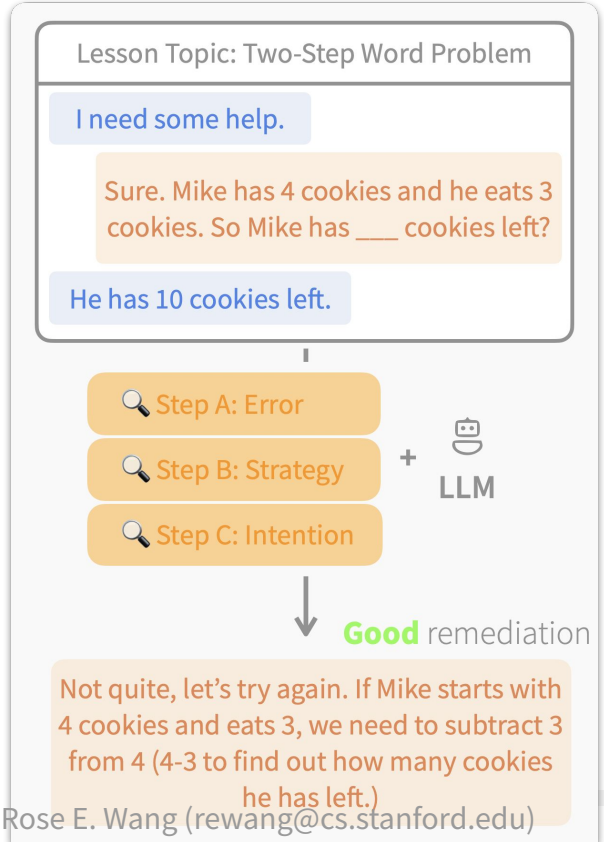
Scaling good pedagogy is hard & expensive!

Experienced teachers **do engage** with the student.

Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Let's surface experienced teachers' *decision process* in remediation with Bridge



Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes

Rose E. Wang Qingyang Zhang Carly Robinson
Susanna Loeb Dorottya Demszky

Scan to read more about this work!

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Rose E. Wang (rewang@cs.stanford.edu)

How can we help novice tutors teach better with real-time decision aids?

The Effect of Tutor CoPilot for Virtual Tutoring Sessions: Testing an Intervention to Improve Tutor Instruction with Expert-Guided LLM-generated Remediation Language

Rose E. Wang, Ana Ribeiro, Carly Robinson, Dorottya Demszky, Susanna Loeb

https://osf.io/8d6ha

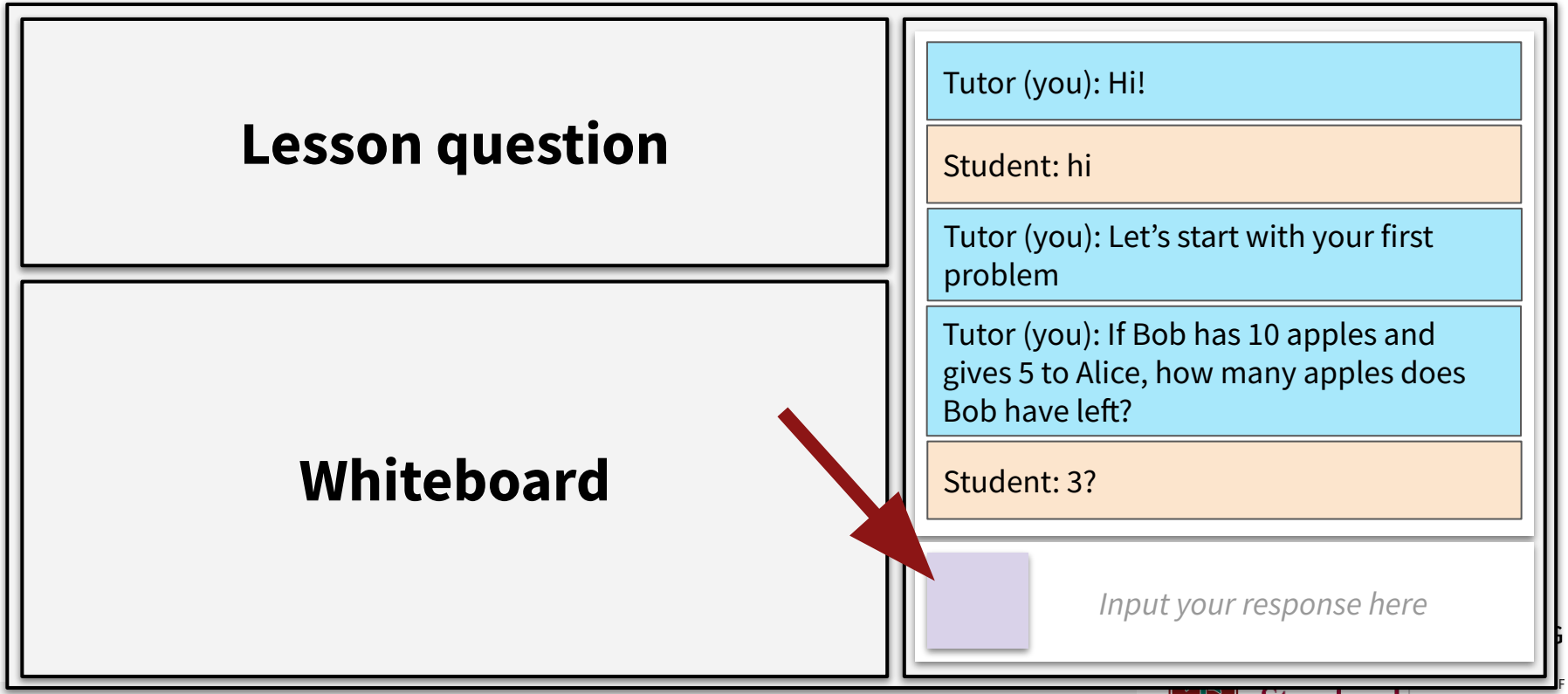




Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

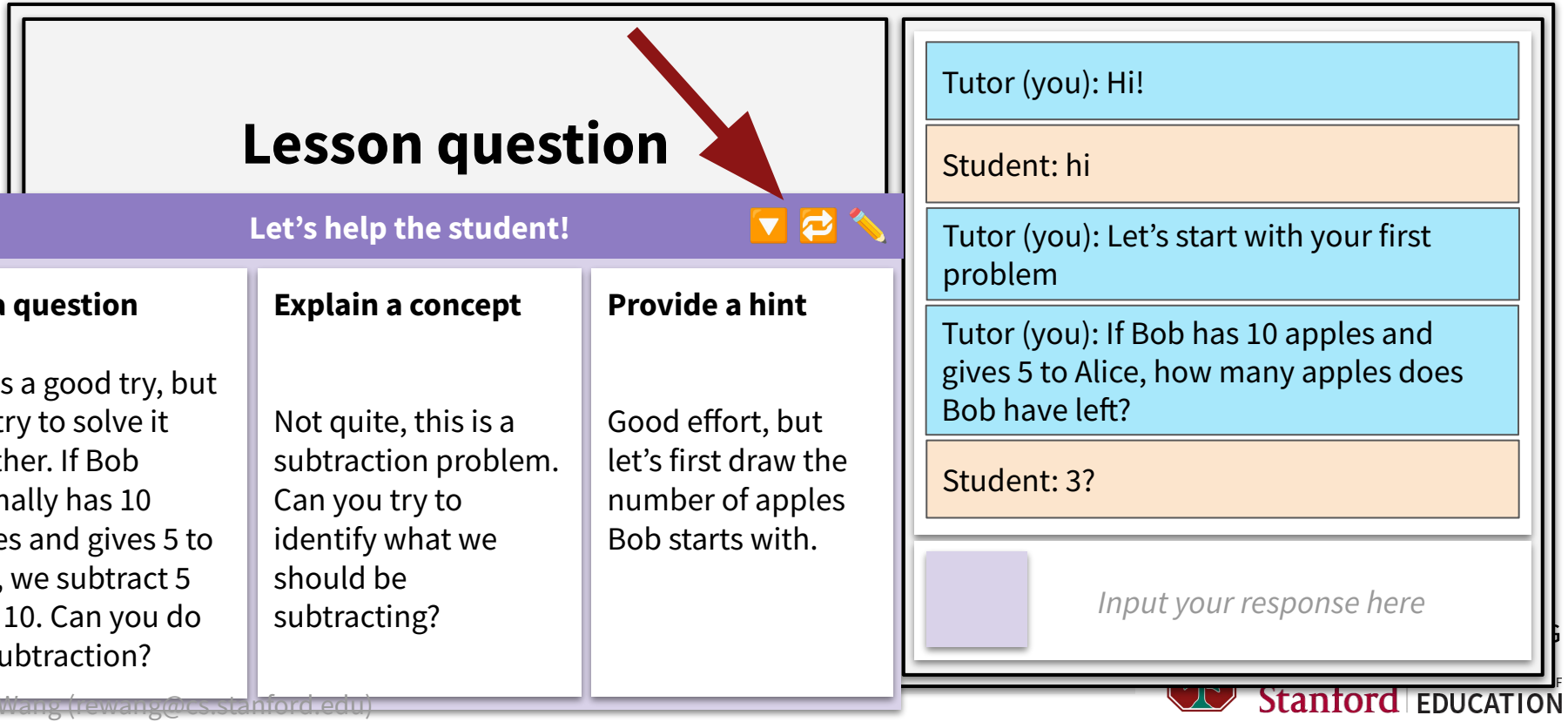
Rose E. Wang (rewang@cs.stanford.edu)

Tutor CoPilot is an AI-in-the-loop, real-time decision aid

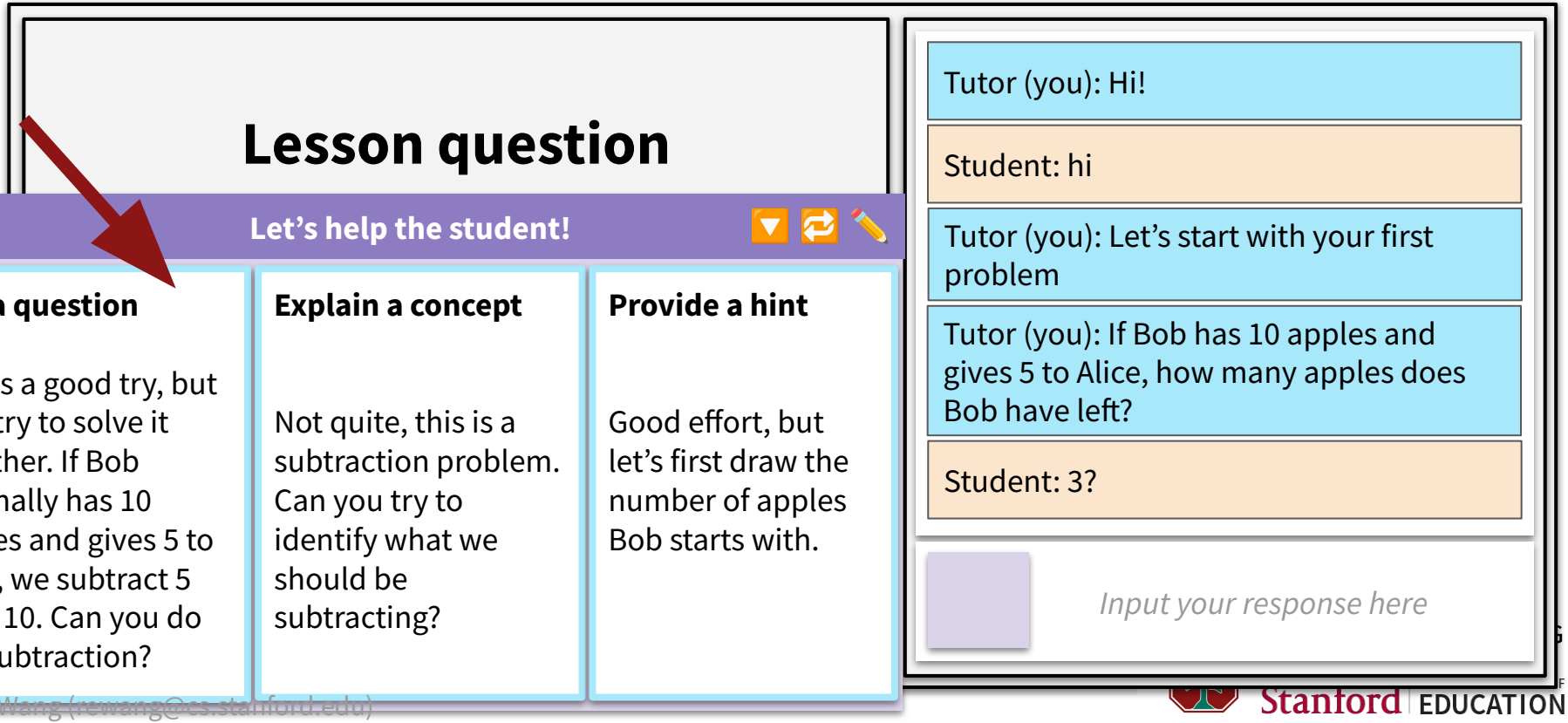


Rose E. Wang (rewang@cs.stanford.edu)

Tutor CoPilot is an AI-in-the-loop, real-time decision aid

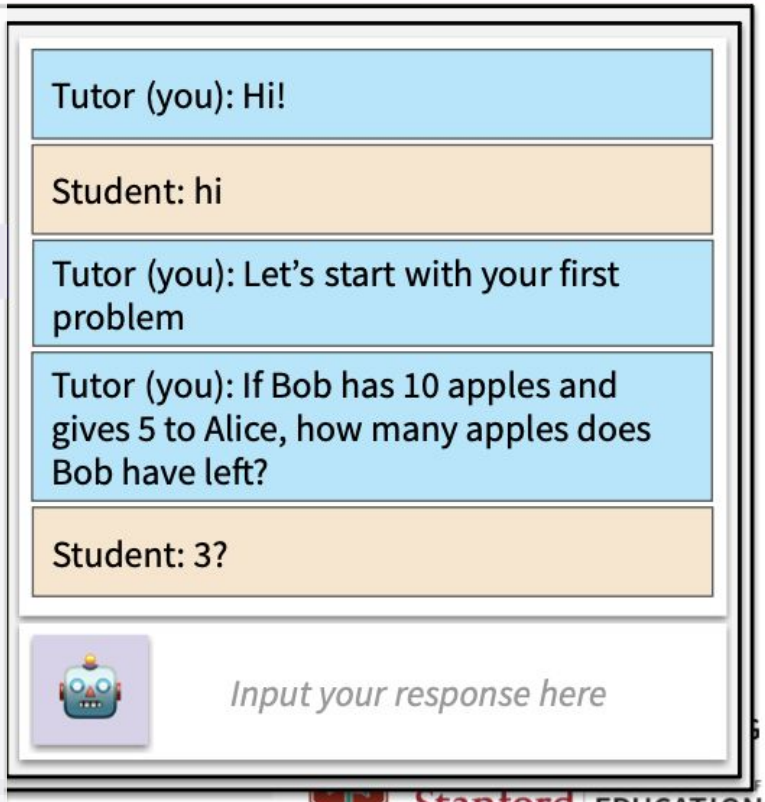


Tutor CoPilot generates suggestions for multiple strategies.



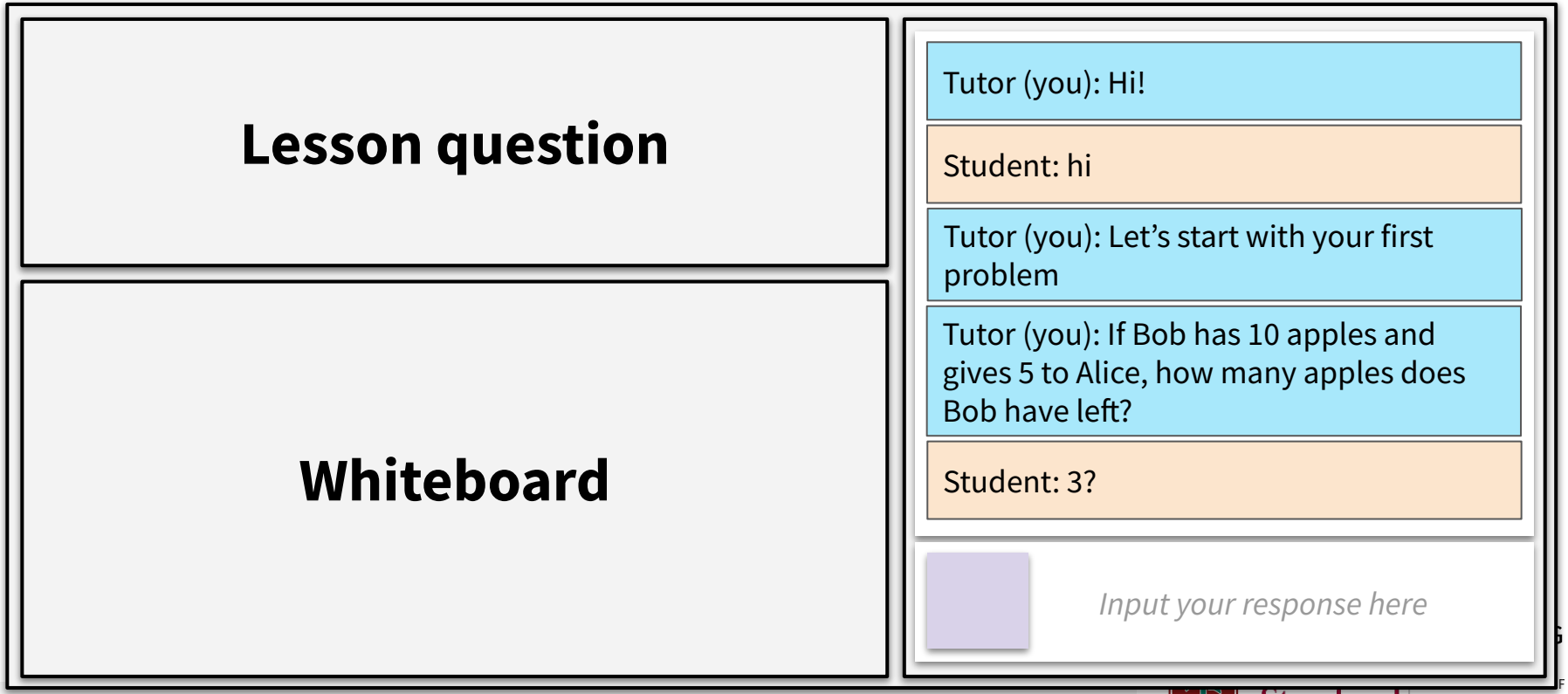
Tutor CoPilot supports the tutor—it doesn't replace.

Tutors possess important knowledge that is not readily apparent from the chat!
e.g., curriculum, previous interactions…



Rose E. Wang (rewang@cs.stanford.edu)

Tutor CoPilot is a tutor-facing real-time decision aid



Rose E. Wang (rewang@cs.stanford.edu)

Sampling plan

Southern school district, 2023-24

In-school, High-dosage tutoring (1-3 sessions/week)

1534 students

874 tutors
- Treatment: 429
- Control: 450

(Numbers not equal because of attrition between initial assignment and study launch)

Rose E. Wang (rewang@cs.stanford.edu)

Randomization

1x2 experimental design. Randomization of **tutors**.

(1) Treatment group: Tutors get access to Tutor CoPilot in their tutoring sessions

(2) Control group: Tutors do not have access and they tutor as normal in their tutoring sessions.

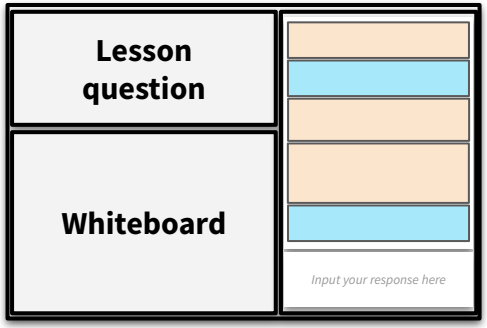
The design plan and outcome variables

Randomized Controlled Trial (1.5 months)

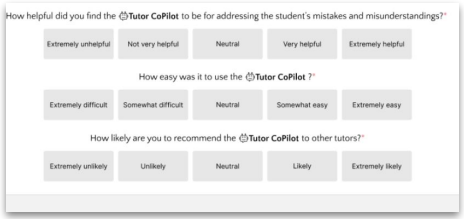
Tutor

Pre-study survey

Tutoring



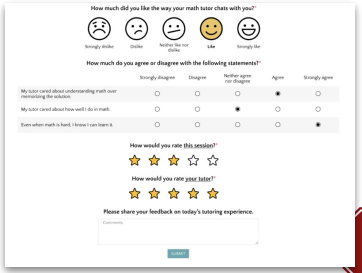
Post tutoring survey



Post-study survey

Student

Winter MAP

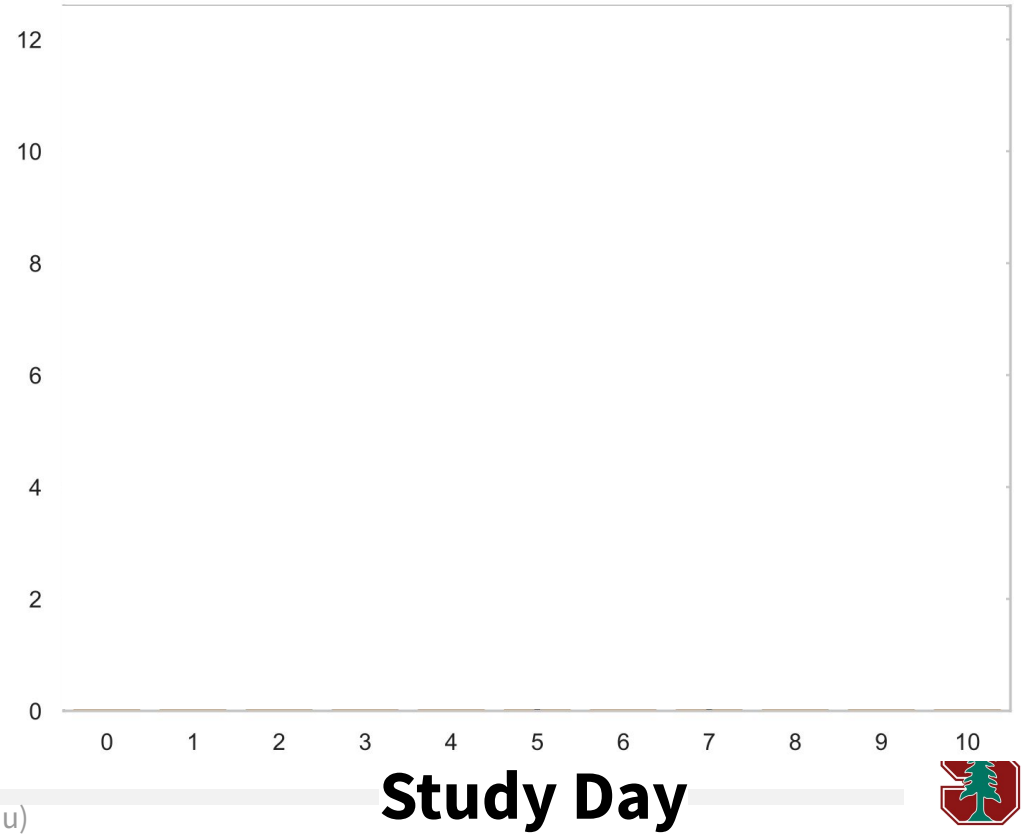


Spring MAP

Rose E. Wang (rewang@cs.stanford.edu)

Treatment Tutors use Tutor CoPilot 2-4 times per session

Number of Tutor CoPilot Uses in Session



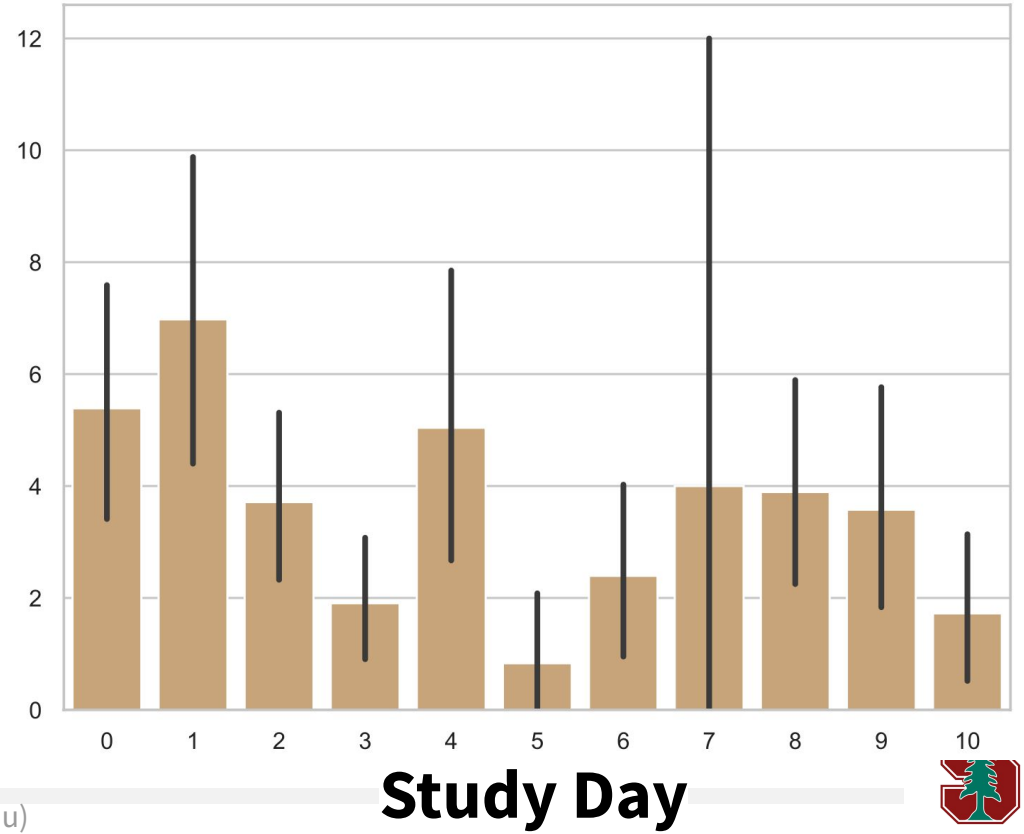
Preliminary findings

Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Treatment Tutors use Tutor CoPilot 2-4 times per session

Number of Tutor CoPilot Uses in Session



Preliminary findings

Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Tutor CoPilot results in positive student perception of the tutor and tutoring session

Preliminary findings

Panel B. Student survey outcomes

Dependent Variable:	My tutor cared about understanding math over memorizing the solution.	My tutor cared about how well I do in math.	Even when math is hard, I know I can learn it.
Treatment Effect			
Control Mean			
n	410	410	410

*Note. +p<0.1; *p<0.05; **p<0.01, ***p<0.001, Estimates are from our primary model, which controls for baseline math scores, student demographics, and a fixed effect for strata (school x grade).*

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Tutor CoPilot results in positive student perception of the tutor and tutoring session

Preliminary findings

Panel B. Student survey outcomes

Dependent Variable:	My tutor cared about understanding math over memorizing the solution.	My tutor cared about how well I do in math.	Even when math is hard, I know I can learn it.
Treatment Effect	0.163 (0.124)	0.261* (0.121)	0.215+ (0.125)
Control Mean	4.122 (0.068)	4.232 (0.067)	4.203 (0.067)
n	410	410	410

*Note. +p<0.1; *p<0.05; **p<0.01, ***p<0.001, Estimates are from our primary model, which controls for baseline math scores, student demographics, and a fixed effect for strata (school x grade).*

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Positive trends in session outcomes like participation points awarded.

Preliminary findings

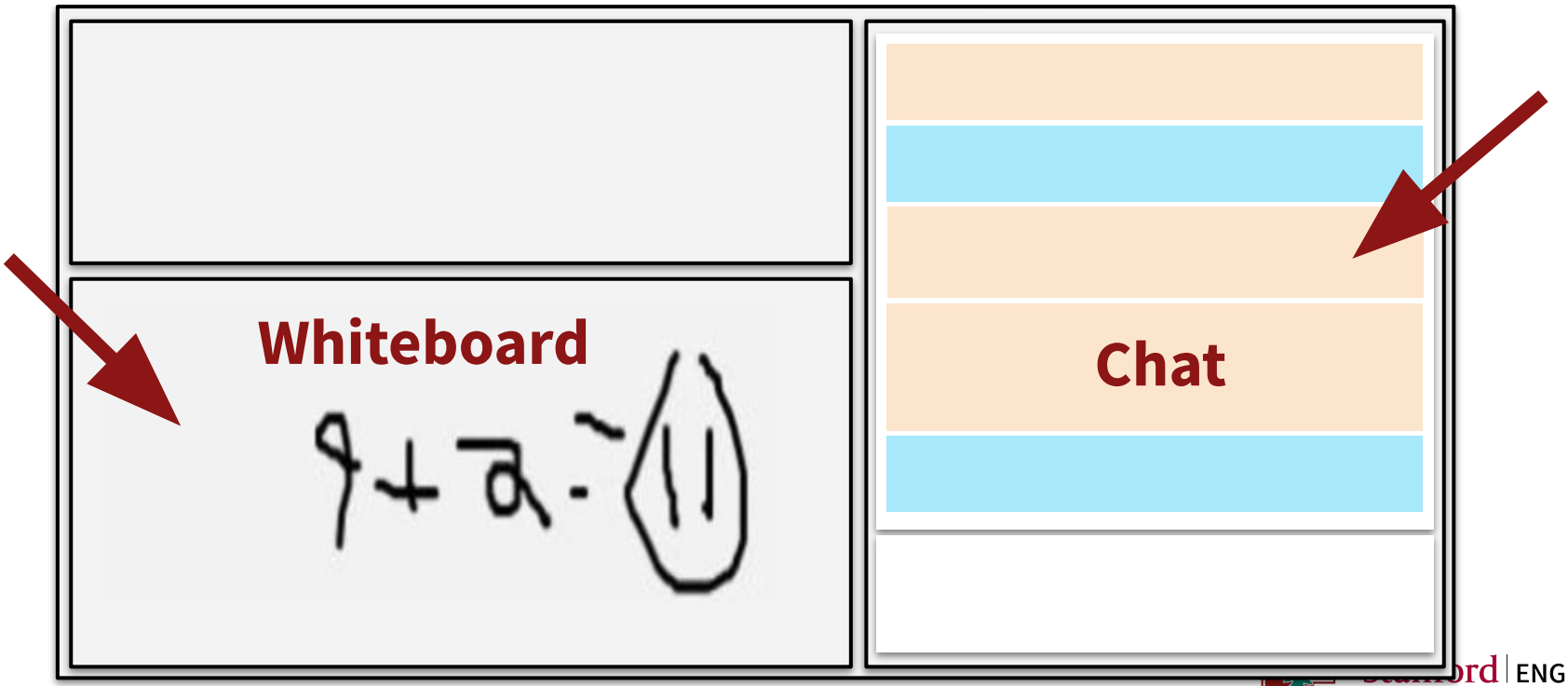
Panel A. Session outcomes and Ratings

Dependent Variable:	Participation Points	Participation Points (Standardized)
Treatment Effect	0.3117 (0.4715)	0.0491 (0.0508)
Control Mean	13.1744 (0.4307)	0.0075 (0.0424)
n	1074	1074

Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

We plan to additionally measure student and tutor engagement via chat and whiteboard.



Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION

Real-time decision aids for improving human decision-making

Teaching is *hard*: Backstage, hidden decisions being made.

Tutor CoPilot **:** An AI-in-the-loop, real-time decision aid for improving tutoring instruction.

Future open questions: e.g., Who are these tools most effective for?

Future use cases: e.g., How can these tools improve other areas of interaction and instruction?

Rose E. Wang (rewang@cs.stanford.edu)

Stanford | ENGINEERING
Computer Science
Stanford | GRADUATE SCHOOL OF EDUCATION