CS 329X: Human Centered LLMs

Enabling Human-AI Interaction

Diyi Yang

Announcements

- Project proposal due, Oct 16th, 2025
- Homework 1
- Homework 2 out (more to see at 5:10pm today!!)

- Human annotation / help each other :D

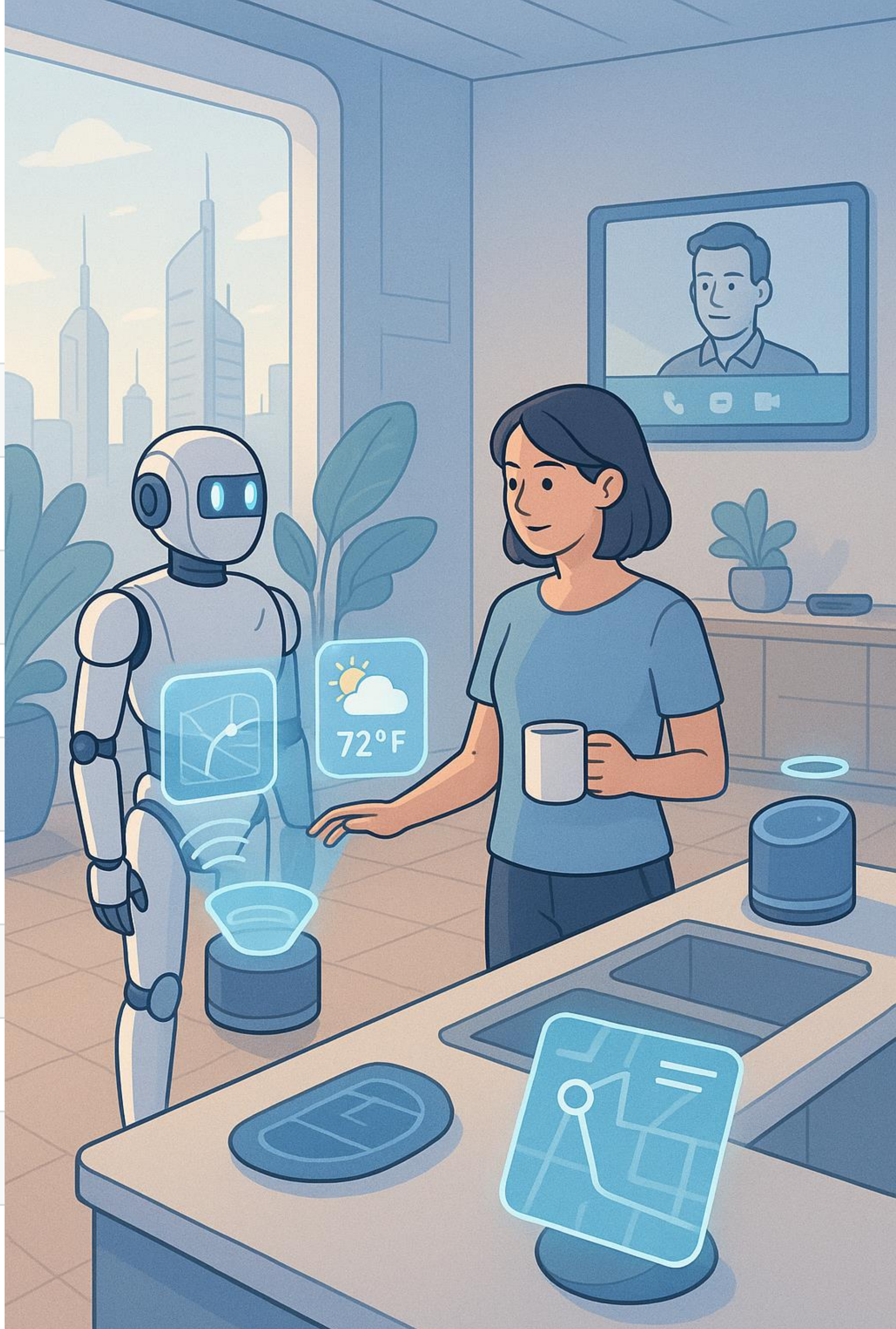
Outline

- ❑ **What Is Human-AI Interaction** (10 mins)
- ❑ **Enable Human-AI Interaction & Collaboration** (20 mins)
- ❑ **Human-AI Interaction Case Studies** (10 mins)
- ❑ **Hot-take Debate** (20 mins)

Learning Objective: understand different types of human-AI interaction, and other approaches to enable human-AI interaction/collaboration

🌍 A Day in the Life with AI

Time	Interaction Example
⏰ Morning	Smart alarm & voice assistant
🚗 Commute	Traffic updates, autonomous driving
🖥 Work	Smart scheduling, copilot
🛒 Shopping	Product recommendations
📚 Education	Personalized learning apps
💰 Finance	Budgeting apps, fraud detection
🏥 Health	Fitness trackers, symptom checkers
💬 Social	Chatbots, real-time translation
🧘 Night	Sleep quality insights, guided meditations



The Rise of “Human-AI Interaction”

Definition: how humans and AI systems interact

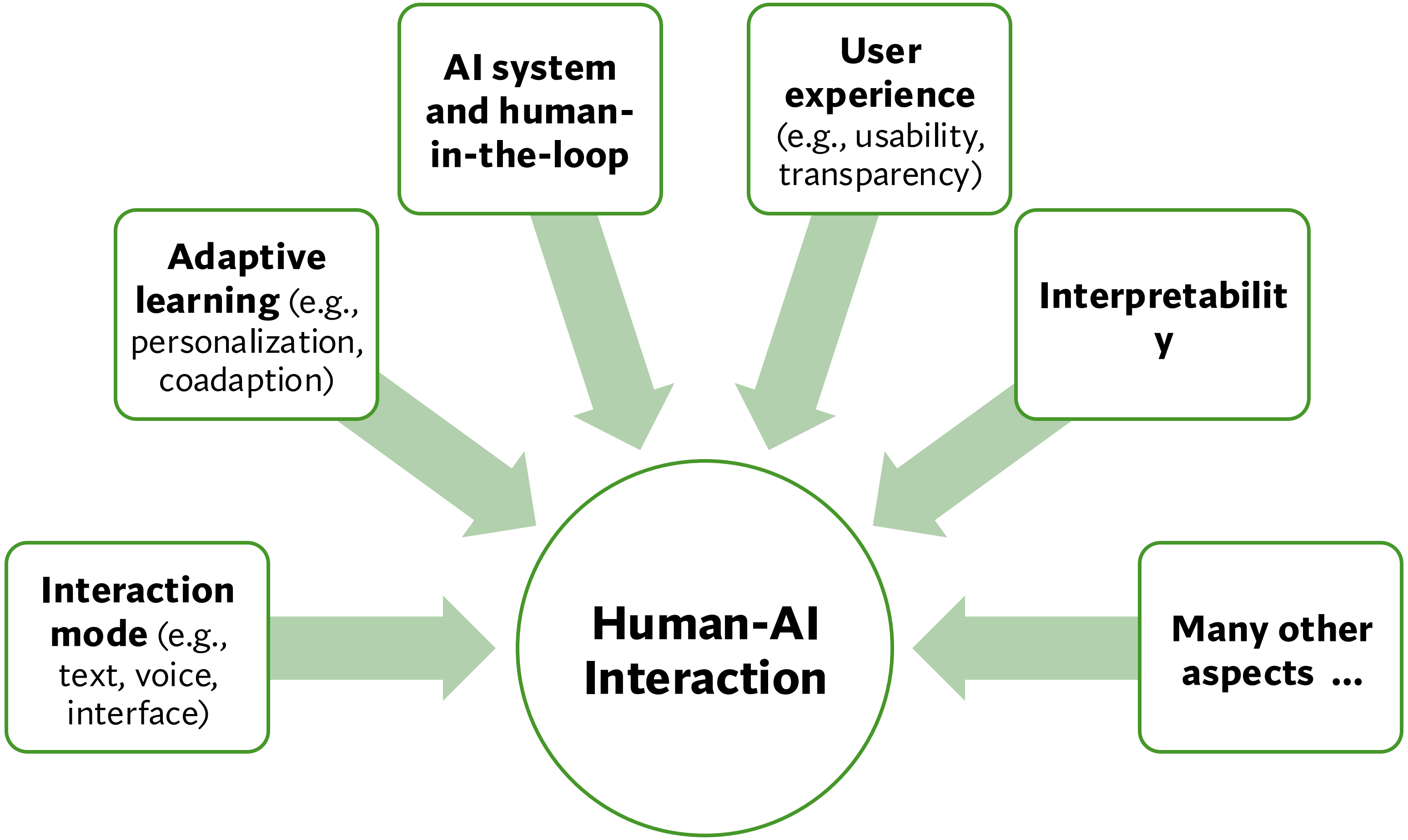
Humans: AI researchers, model developers, domain experts, end users...

AIs: LLMs, VLMs, translator, recommender system, autonomous driving system...

Interact:

Humans collaborate with AI,
Humans get assistance from AI,
Humans analyze AI,
AI helps human,
& many other forms

The Space of Human-AI Interaction



Human-AI Interaction

Interaction mode (e.g., text, voice, interface)

Adaptive learning (e.g., personalization, coadaption)

AI system and human-in-the-loop

User experience (e.g., usability, transparency)

Interpretabilit y

Many other aspects …

Theoretical foundation: HCI, psychology, sociotechnical systems, design, ethics, cognitive science

How Should We Build Better Human-AI Interaction

(self-)selected *Already exist <u>but needs improvement!</u>*

Given a **human** and an **AI/LLM...**

Design Why should they interact? How do we make it happen?

Enable How can we enable human-LLM interaction?

Evaluate Have we achieved what we want to achieve?

Outline

✓**What Is Human-AI Interaction** (10 mins)

❑**Enable Human-AI Interaction & Collaboration** (20 mins)

- ❑Automation vs. Augmentation in “Human-AI Collaboration”
- ❑Agency, RL and situational reasoning to improve collaboration
- ❑Mixed examples of “does human-AI collaboration work”

How Should We Build Better Human-AI Interaction

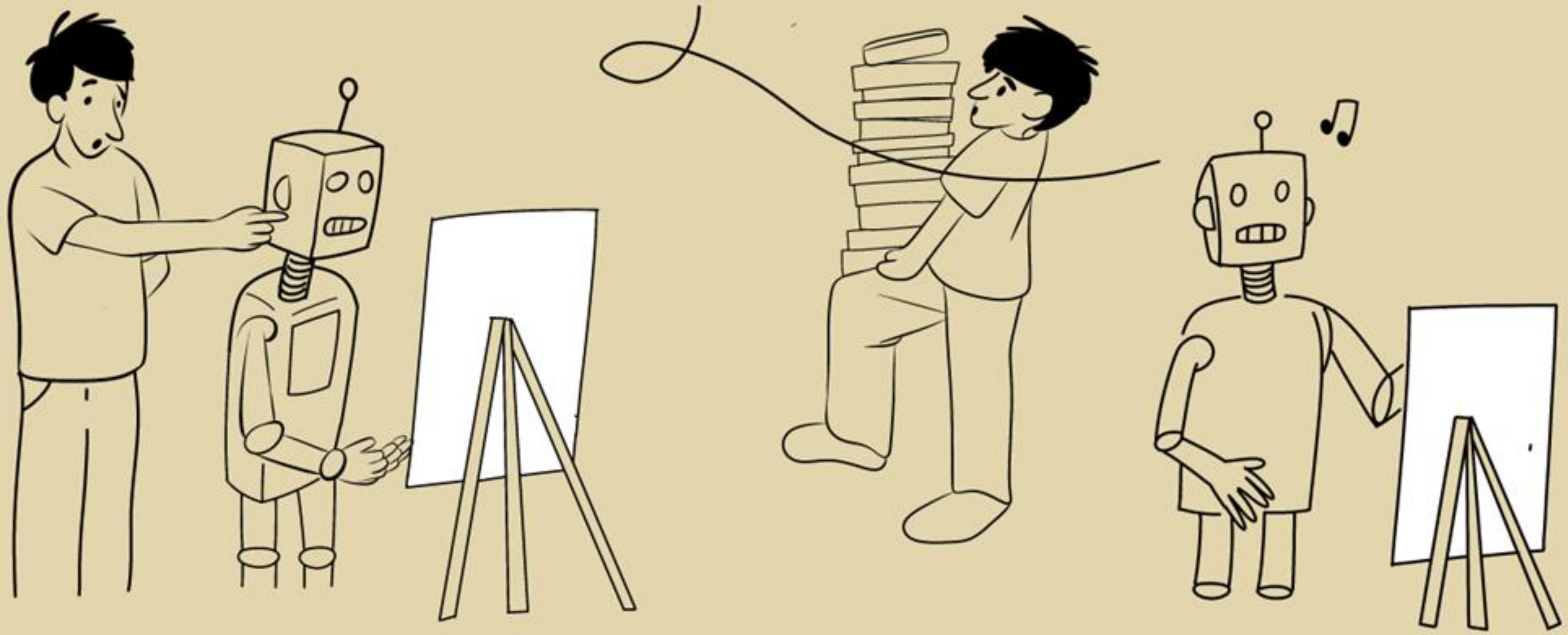
(self-)selected *Already exist but needs improvement!*

Given a **human** and an **AI/LLM…**

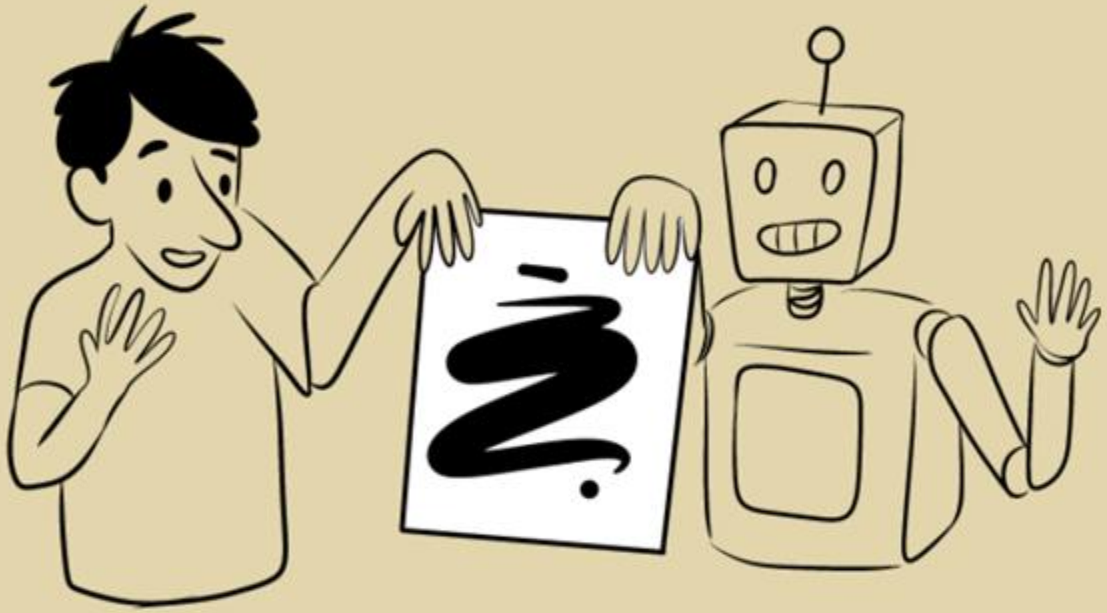
Design Why should they interact? How do we make it happen?

Enable How can we enable human-LLM interaction?

Evaluate Have we achieved what we want to achieve?



AI was supposed to take boring jobs and allow people to be creative, **but the opposite happened**



Outline

✓**What Is Human-AI Interaction** (10 mins)

❑**Enable Human-AI Interaction & Collaboration** (20 mins)

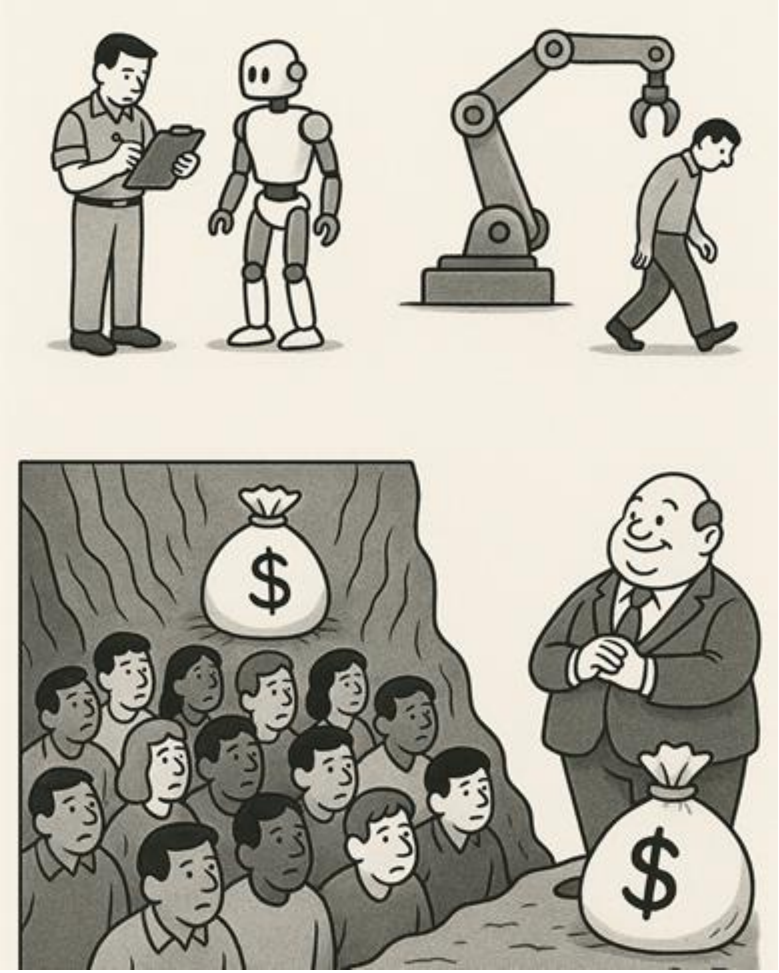
- ❑Automation vs. Augmentation in “Human-AI Collaboration”
- ❑Agency, RL and situational reasoning to improve collaboration
- ❑Mixed examples of “does human-AI collaboration work”

Automation vs. Augmentation

Automation: AI replaces human capabilities

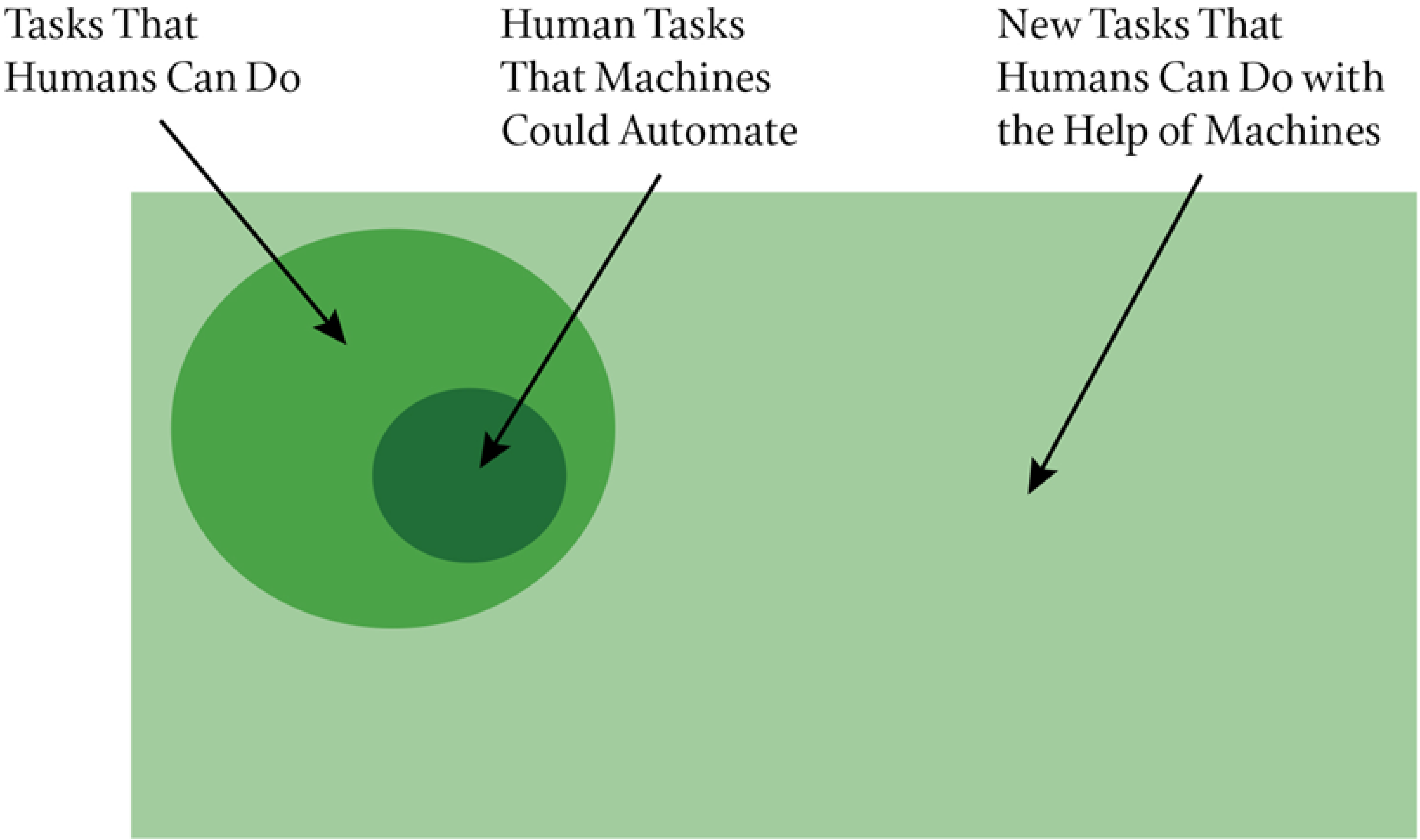
Augmentation: AI enhances human capabilities

- ❖ Automation isn’t inherently bad
- ❖ Automation increases profits for a few
- ❖ Augmentation can lead to increased performance, creativity, new tasks, and emerging needs



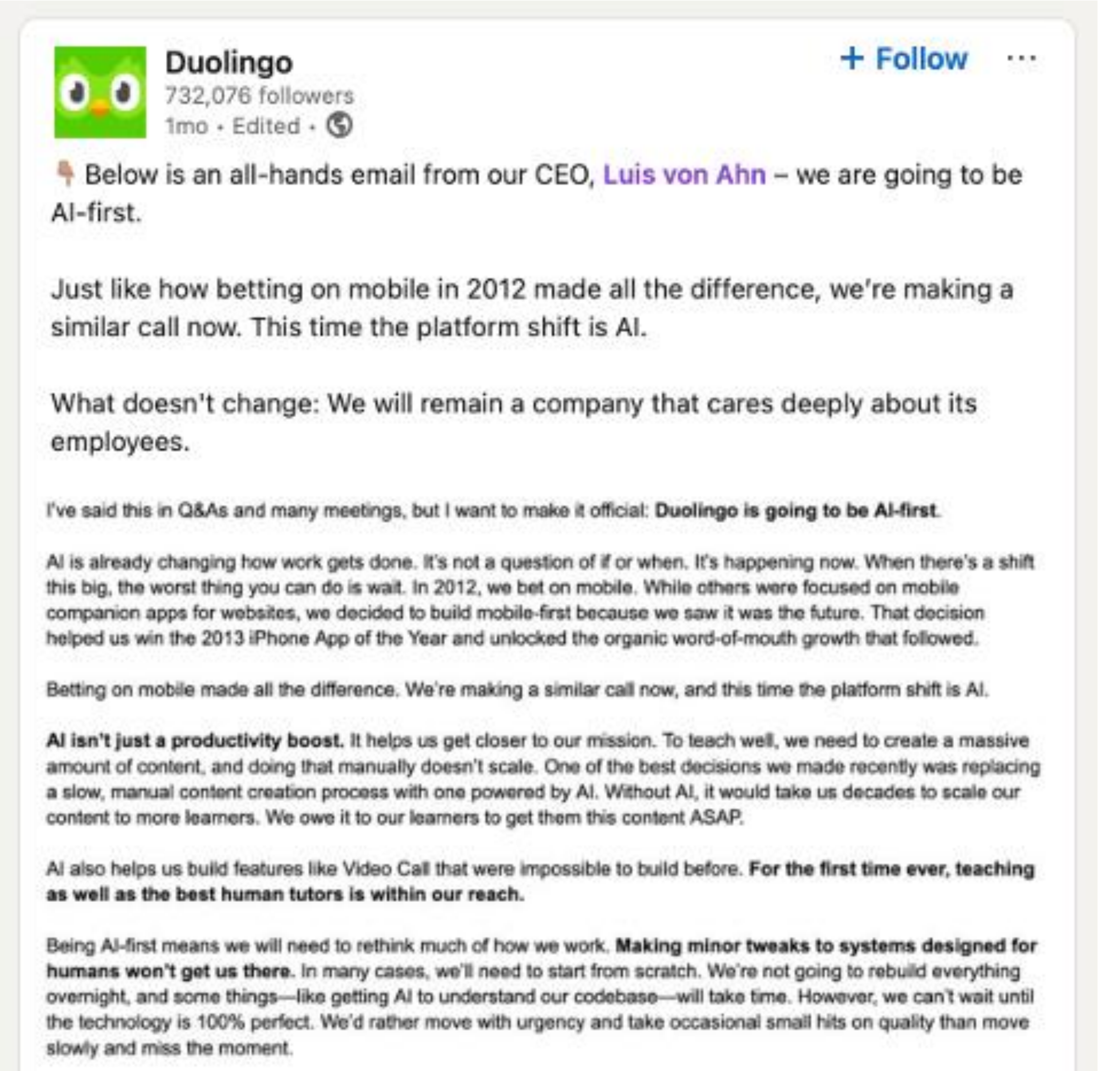
Turing Trap

Automating work instead of enhancing human labor leads to a concentration of wealth and power, trapping the rest in a system where they can’t improve their lives


Tasks That Humans Can Do
Human Tasks That Machines Could Automate
New Tasks That Humans Can Do with the Help of Machines

Brynjolfsson, Erik. "The turing trap: The promise & peril of human-like artificial intelligence." *Daedalus* 151, no. 2 (2022): 272-287.

Recent Examples on Automation vs. Augmentation



👇 Below is an all-hands email from our CEO, Luis von Ahn – we are going to be AI-first.

Just like how betting on mobile in 2012 made all the difference, we're making a similar call now. This time the platform shift is AI.

What doesn't change: We will remain a company that cares deeply about its employees.

I've said this in Q&As and many meetings, but I want to make it official: **Duolingo is going to be AI-first.**

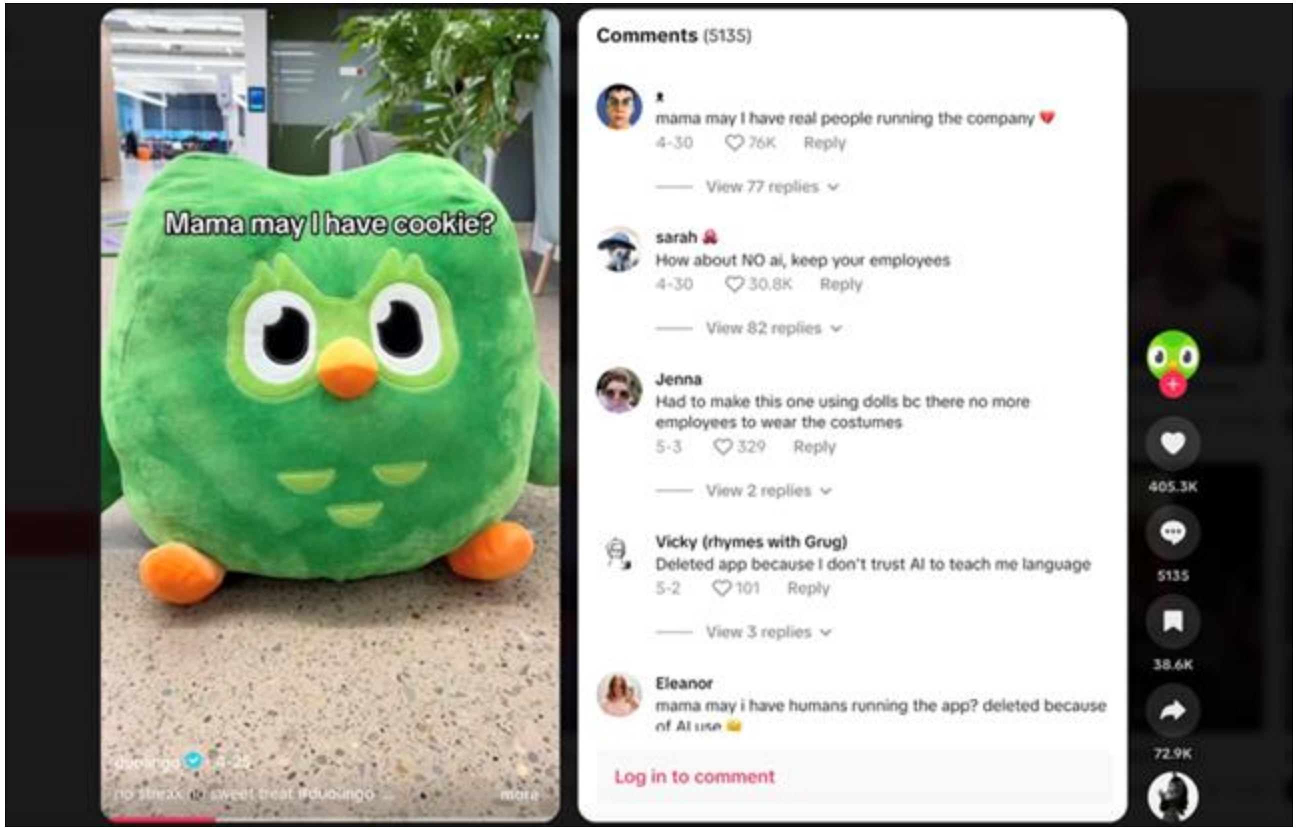
AI is already changing how work gets done. It's not a question of if or when. It's happening now. When there's a shift this big, the worst thing you can do is wait. In 2012, we bet on mobile. While others were focused on mobile companion apps for websites, we decided to build mobile-first because we saw it was the future. That decision helped us win the 2013 iPhone App of the Year and unlocked the organic word-of-mouth growth that followed.

Betting on mobile made all the difference. We're making a similar call now, and this time the platform shift is AI.

AI isn't just a productivity boost. It helps us get closer to our mission. To teach well, we need to create a massive amount of content, and doing that manually doesn't scale. One of the best decisions we made recently was replacing a slow, manual content creation process with one powered by AI. Without AI, it would take us decades to scale our content to more learners. We owe it to our learners to get them this content ASAP.

AI also helps us build features like Video Call that were impossible to build before. **For the first time ever, teaching as well as the best human tutors is within our reach.**

Being AI-first means we will need to rethink much of how we work. **Making minor tweaks to systems designed for humans won't get us there.** In many cases, we'll need to start from scratch. We're not going to rebuild everything overnight, and some things—like getting AI to understand our codebase—will take time. However, we can't wait until the technology is 100% perfect. We'd rather move with urgency and take occasional small hits on quality than move slowly and miss the moment.



Comments (5135)

mama may I have real people running the company ❤️
4-30 ♡ 76K Reply
View 77 replies

sarah
How about NO ai, keep your employees
4-30 ♡ 30.8K Reply
View 82 replies

Jenna
Had to make this one using dolls bc there no more employees to wear the costumes
5-3 ♡ 329 Reply
View 2 replies

Vicky (rhymes with Grug)
Deleted app because I don't trust AI to teach me language
5-2 ♡ 101 Reply
View 3 replies

Eleanor
mama may i have humans running the app? deleted because of AI use

Log in to comment

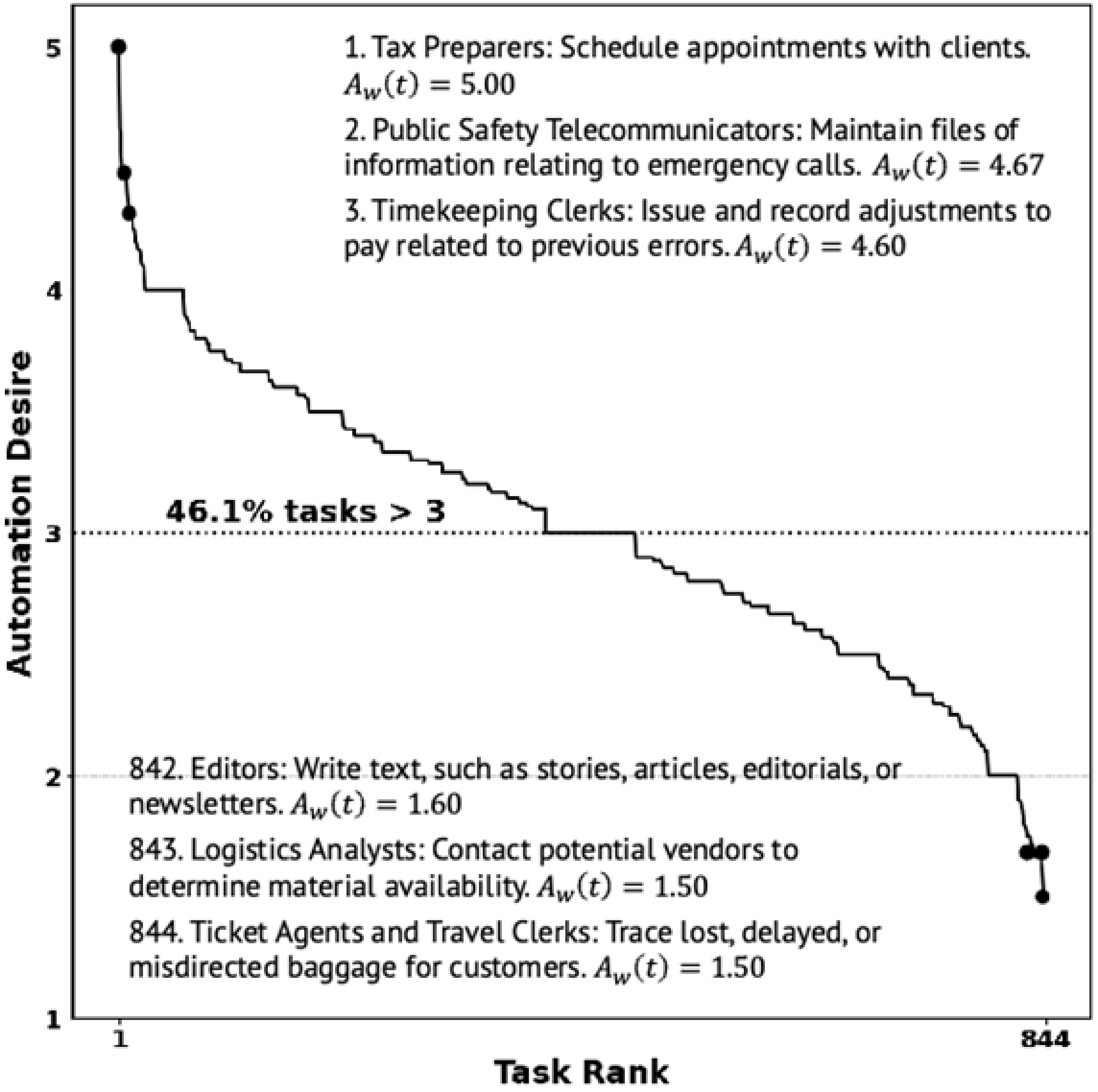
Framework for Automation vs. Augmentation

Human Agency Scale

	HAS H1	HAS H2	HAS H3	HAS H4	HAS H5
Team Dynamics	**AI Agent Drives Task Completion** The AI agent takes primary responsibility for task execution with no or minimal human oversight.		**Equal Partnership** The human and the AI agent collaborate closely throughout the task.	**Human Drives Task Completion** The human takes primary responsibility for task execution with varying levels of AI assistance.	
Required Human Involvement	AI agent handles the task entirely on its own without your involvement.	AI agent needs your input at a few key points to achieve better task performance.	AI agent and you work together to outperform either alone.	AI agent needs your input to successfully complete the task.	Task completion fully relies on your involvement.
AI Role	**Automation** AI replaces human capabilities		**Augmentation** AI enhances human capabilities		

Shao, Yijia, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang.
"Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the US Workforce." arXiv:2506.

Worker Desire for Automation



Top reasons of why workers want automation for these tasks:

1. Automating the task would free up my time for high-value work
2. Automating this task would improve the quality of work
3. The task is mentally draining
4. The task is complicated or difficult

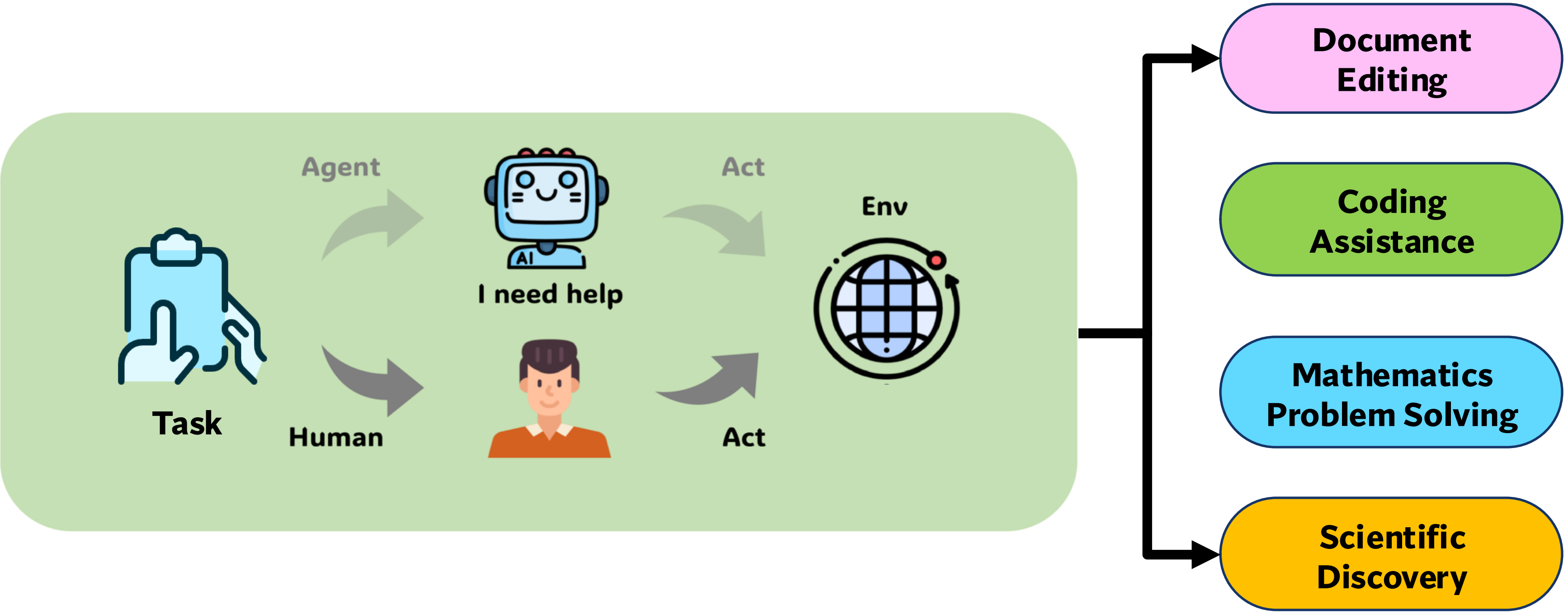
Outline

✓**What Is Human-AI Interaction** (10 mins)

❑**Enable Human-AI Interaction & Collaboration** (20 mins)

- ✓Automation vs. Augmentation in “Human-AI Collaboration”
- ❑Agency, RL and situational reasoning to improve collaboration
- ❑Mixed examples of “does human-AI collaboration work”

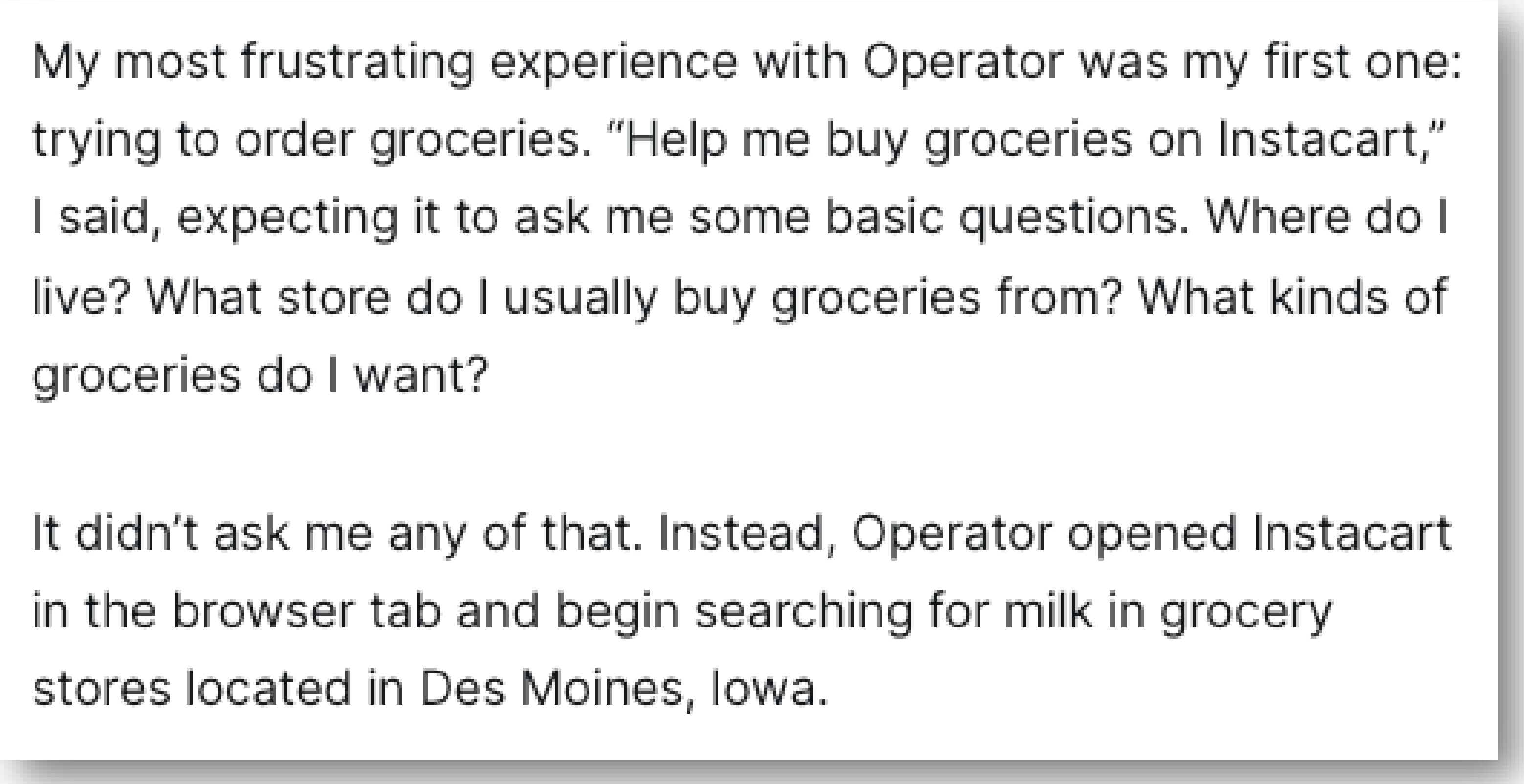
Lots of Applications of Human-AI Interaction/ Collaboration



Slides credit to Shirley Wu

Agency and Friction during Human-AI Interaction

JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY
COMING SOON

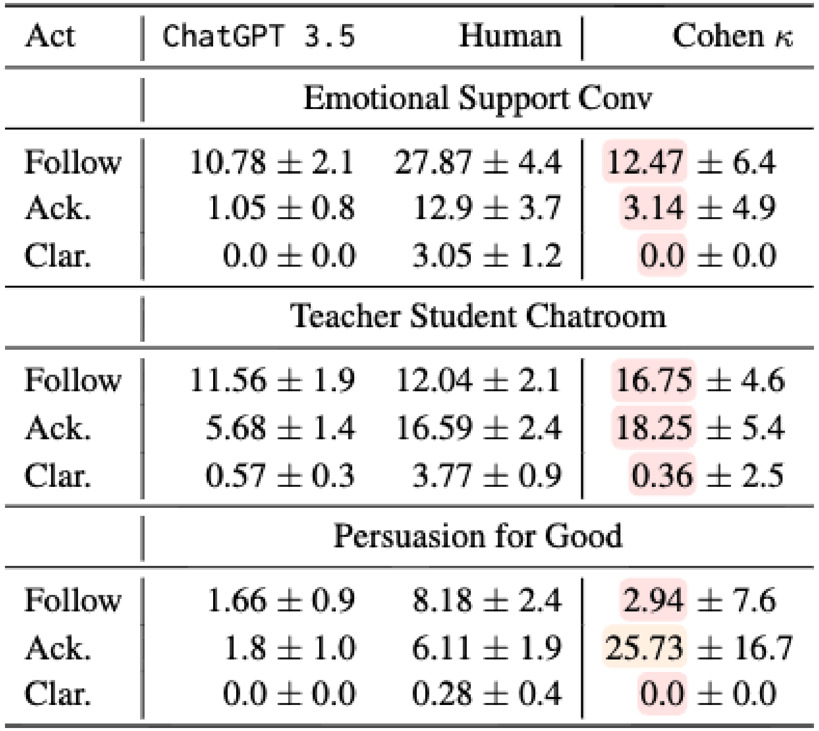


My most frustrating experience with Operator was my first one: trying to order groceries. “Help me buy groceries on Instacart,” I said, expecting it to ask me some basic questions. Where do I live? What store do I usually buy groceries from? What kinds of groceries do I want?

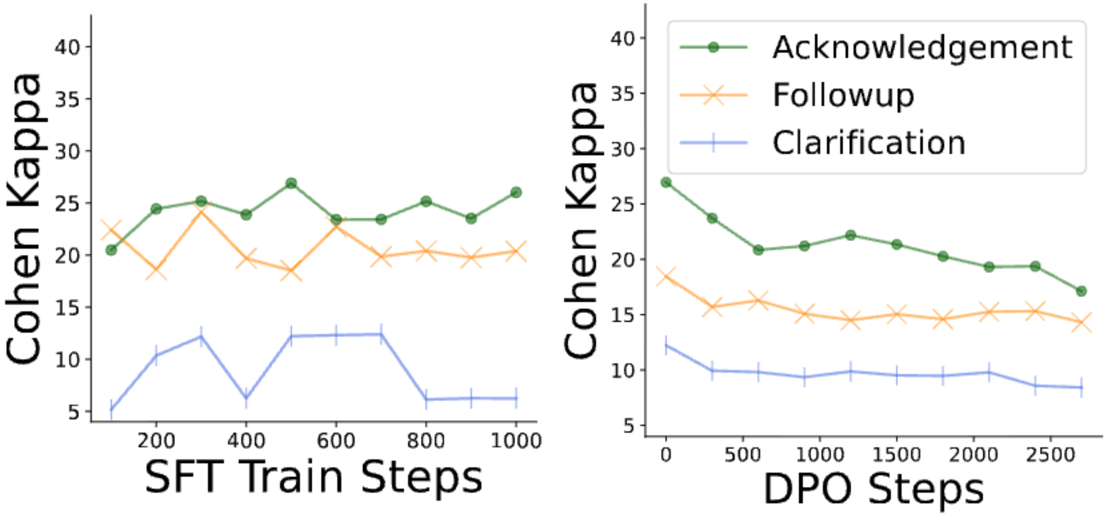
It didn’t ask me any of that. Instead, Operator opened Instacart in the browser tab and begin searching for milk in grocery stores located in Des Moines, Iowa.

https://www.platformer.news/openai-operator-ai-agent-hands-on/

Grounding Gaps in Human-AI Interaction



Act	ChatGPT 3.5	Human	Cohen κ
	Emotional Support Conv		
Follow	10.78 ± 2.1	27.87 ± 4.4	12.47 ± 6.4
Ack.	1.05 ± 0.8	12.9 ± 3.7	3.14 ± 4.9
Clar.	0.0 ± 0.0	3.05 ± 1.2	0.0 ± 0.0
	Teacher Student Chatroom		
Follow	11.56 ± 1.9	12.04 ± 2.1	16.75 ± 4.6
Ack.	5.68 ± 1.4	16.59 ± 2.4	18.25 ± 5.4
Clar.	0.57 ± 0.3	3.77 ± 0.9	0.36 ± 2.5
	Persuasion for Good		
Follow	1.66 ± 0.9	8.18 ± 2.4	2.94 ± 7.6
Ack.	1.8 ± 1.0	6.11 ± 1.9	25.73 ± 16.7
Clar.	0.0 ± 0.0	0.28 ± 0.4	0.0 ± 0.0



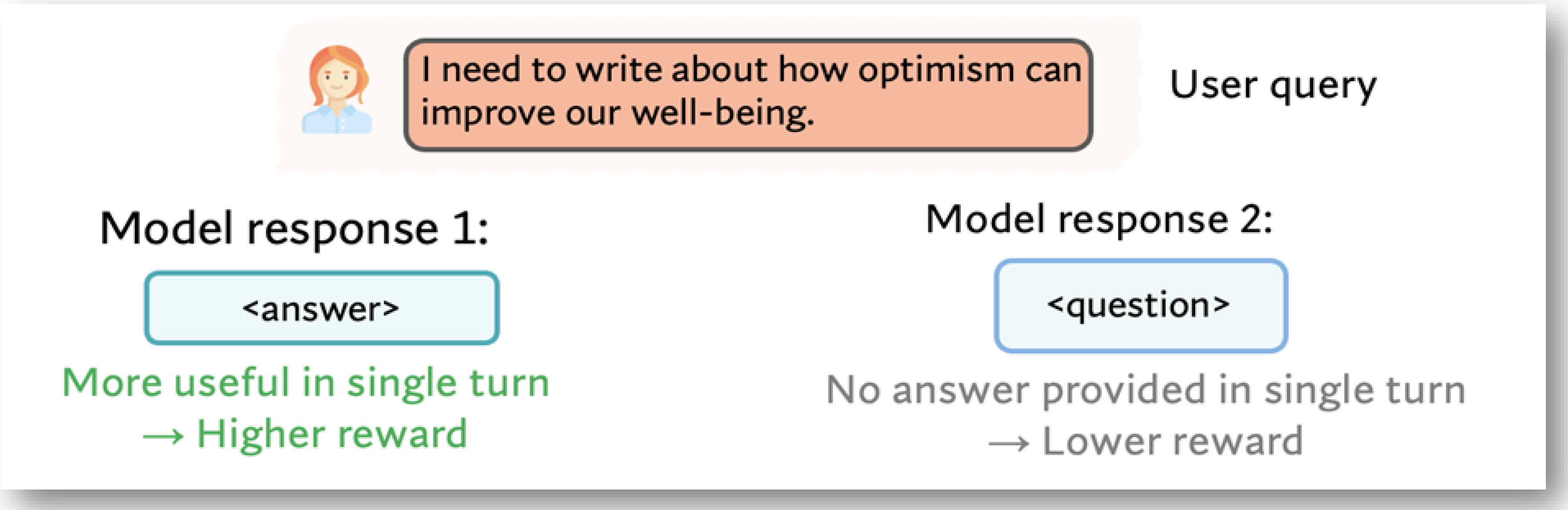
We observe no correlation between SFT training steps and grounding acts, but negative correlation between DPO train steps and grounding acts

Shaikh, Omar, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. "Grounding gaps in language model generations." arXiv preprint arXiv:2311.09144 (2023).

Using RL to Improve Human-AI Collaboration

LLMs are usually tuned based on single-turn human preferences

But these single-turn rewards encourage models to generate responses that may NOT be useful in the long term.



I need to write about how optimism can improve our well-being.

User query

Model response 1:

<answer>

More useful in single turn
→ Higher reward

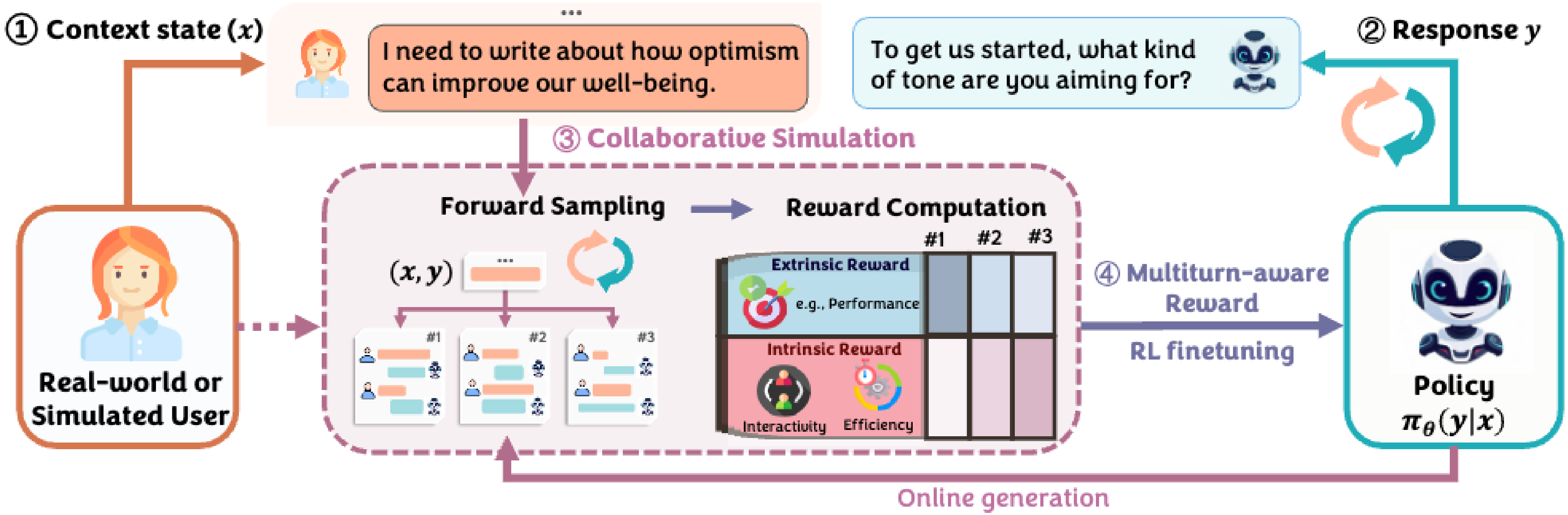
Model response 2:

<question>

No answer provided in single turn
→ Lower reward

Slides credit to Shirley Wu on CollabLLM

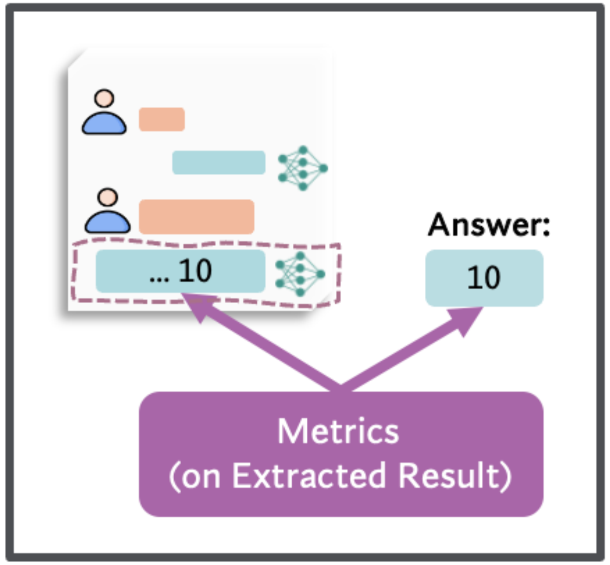
From Passive Responders to Active Collaborators



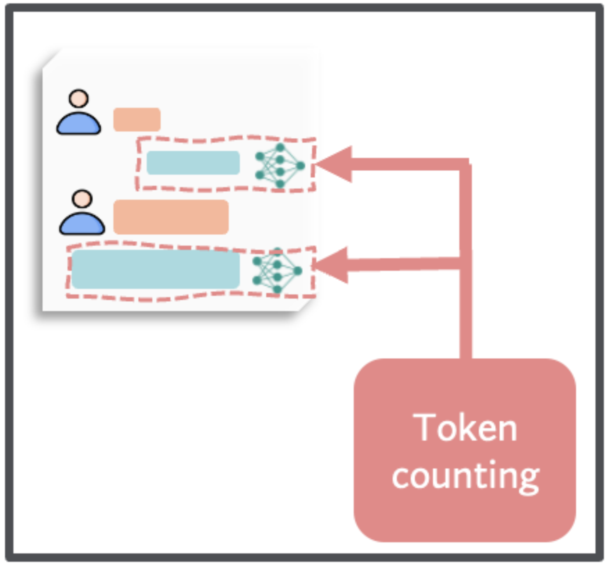
Wu, Shirley, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. "Collabllm: From passive responders to active collaborators." arXiv preprint arXiv:2502.00640 (2025).

Collaborative LLM Reward

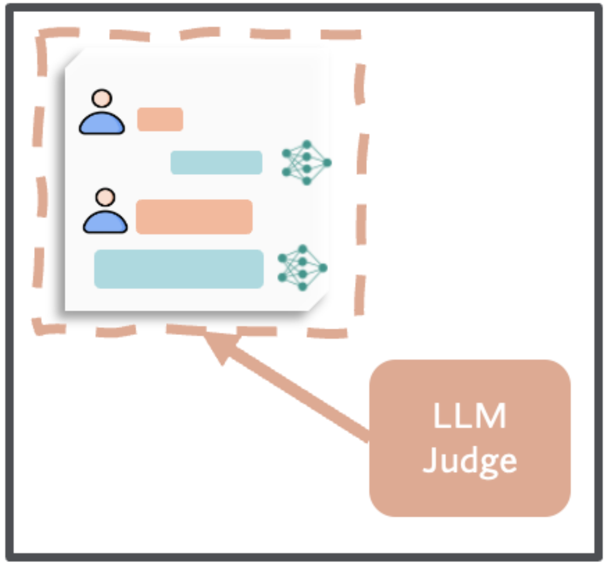
- **Extrinsic Reward:** how well the conversation achieves user's goal
- **Intrinsic Reward:** rewards user experience: token count, LLM judge
- $\boldsymbol{R(t_{1:K})}$**:** Extrinsic/Intrinsic reward or a combination of them



Extrinsic: Task performance
(e.g., accuracy for QA)



Intrinsic:
Efficiency



Intrinsic:
Interactivity

Collaborative LLM Reward

Multiturn-aware reward (MR): Causal effect estimation of how model's response influences the future trajectory of a conversation

Quantify the effect of intervention (model response m_j):

$$MR(m_j) = \mathrm{E}_{t_{1:K} \sim P(t_{1:K} | t_{1:j-1} \cup m_j)} R(t_{1:K})$$

$t_{i:j}$ denotes the conversation from i-th to j-th turn

Estimate long-term impact with synthetic conversations

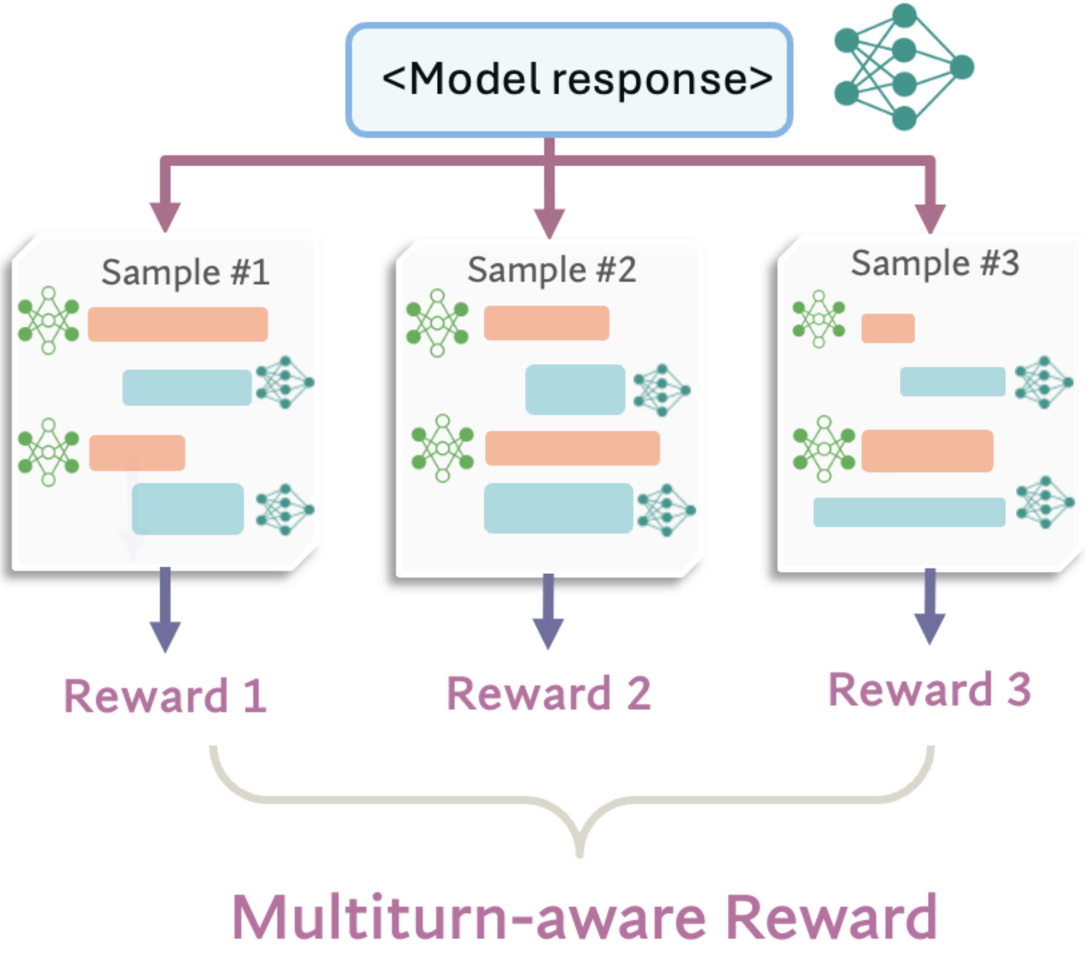
Goal: Evaluate the long-term impact of this model response

Approach:

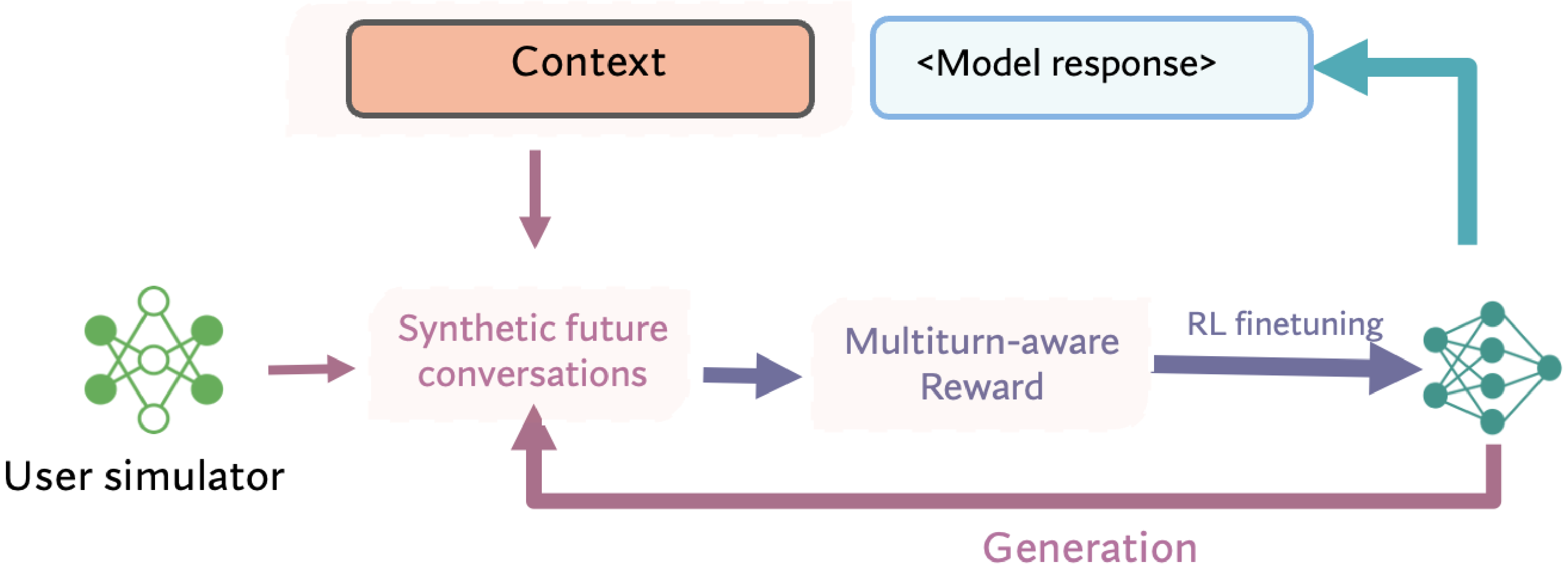
① Sample synthetic future conversations w/ user simulators

② Apply conversation-level reward for each conversation

③ Average the rewards

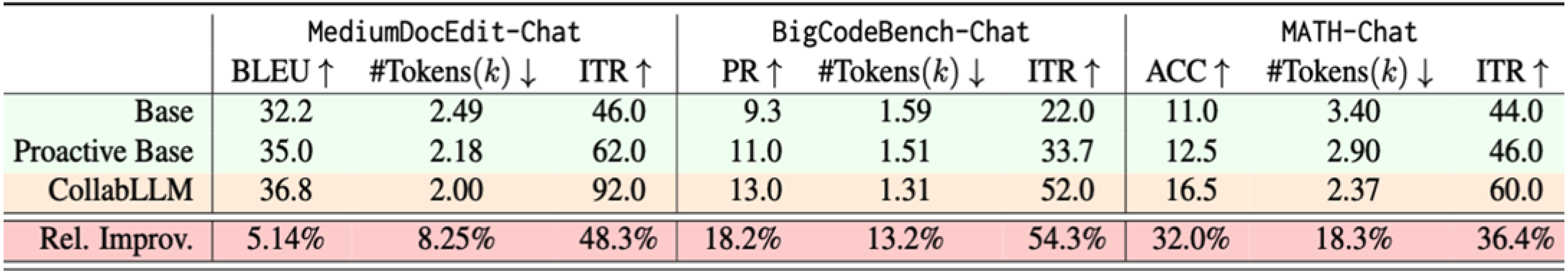


Train LLMs to generate responses that maximize multiturn reward

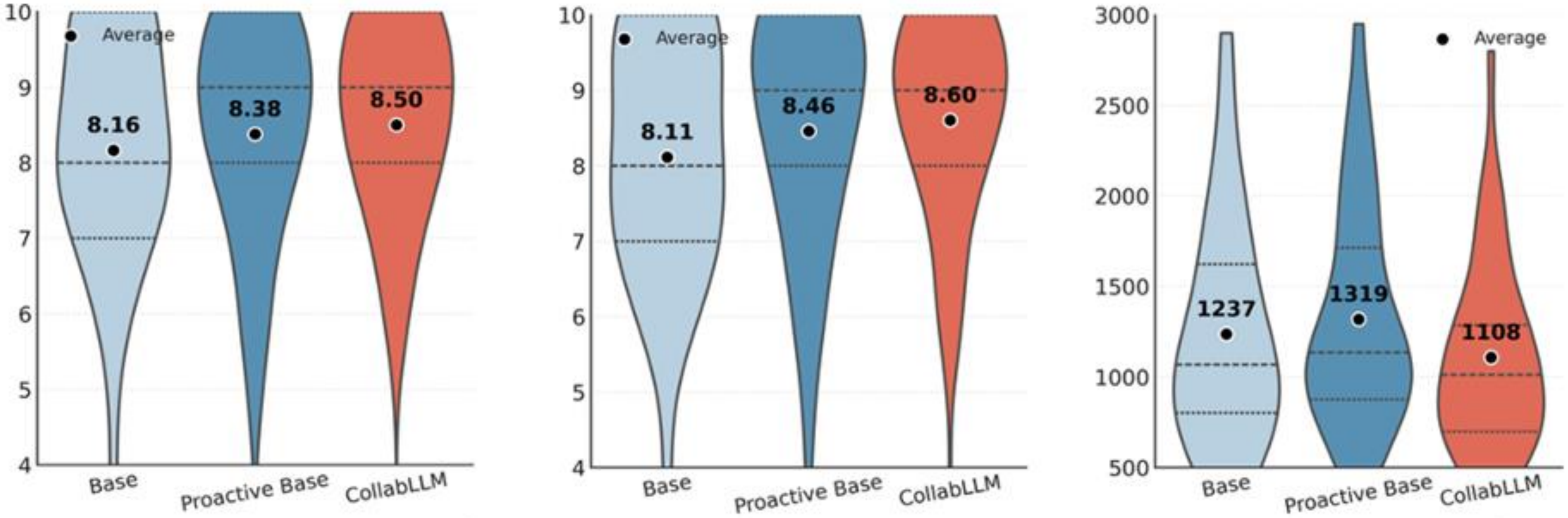


CollabLLM can be integrated with PPO (Schulman et al. 2017) and DPO (Rafailov et al. 2023) to conduct RL finetuning

Improvement of CollabLLM



	MediumDocEdit-Chat			BigCodeBench-Chat			MATH-Chat		
	BLEU ↑	#Tokens(k) ↓	ITR ↑	PR ↑	#Tokens(k) ↓	ITR ↑	ACC ↑	#Tokens(k) ↓	ITR ↑
Base	32.2	2.49	46.0	9.3	1.59	22.0	11.0	3.40	44.0
Proactive Base	35.0	2.18	62.0	11.0	1.51	33.7	12.5	2.90	46.0
CollabLLM	36.8	2.00	92.0	13.0	1.31	52.0	16.5	2.37	60.0
Rel. Improv.	5.14%	8.25%	48.3%	18.2%	13.2%	54.3%	32.0%	18.3%	36.4%



(a) Document Quality Rating ↑

(b) Interaction Rating ↑

(c) Time Spent (s) ↓

CollabLLM yields high-quality documents, better user experience, and saves time by >10%!

Qualitative Insights related to CollaborativeLLMs

About Base (llama-3-1-8b):

“the AI just agreed with me on pretty much everything. There was no debate or discussion.”

About Proactive Base:

”The AI seemed to be very redundant and asked me the same questions over and over”

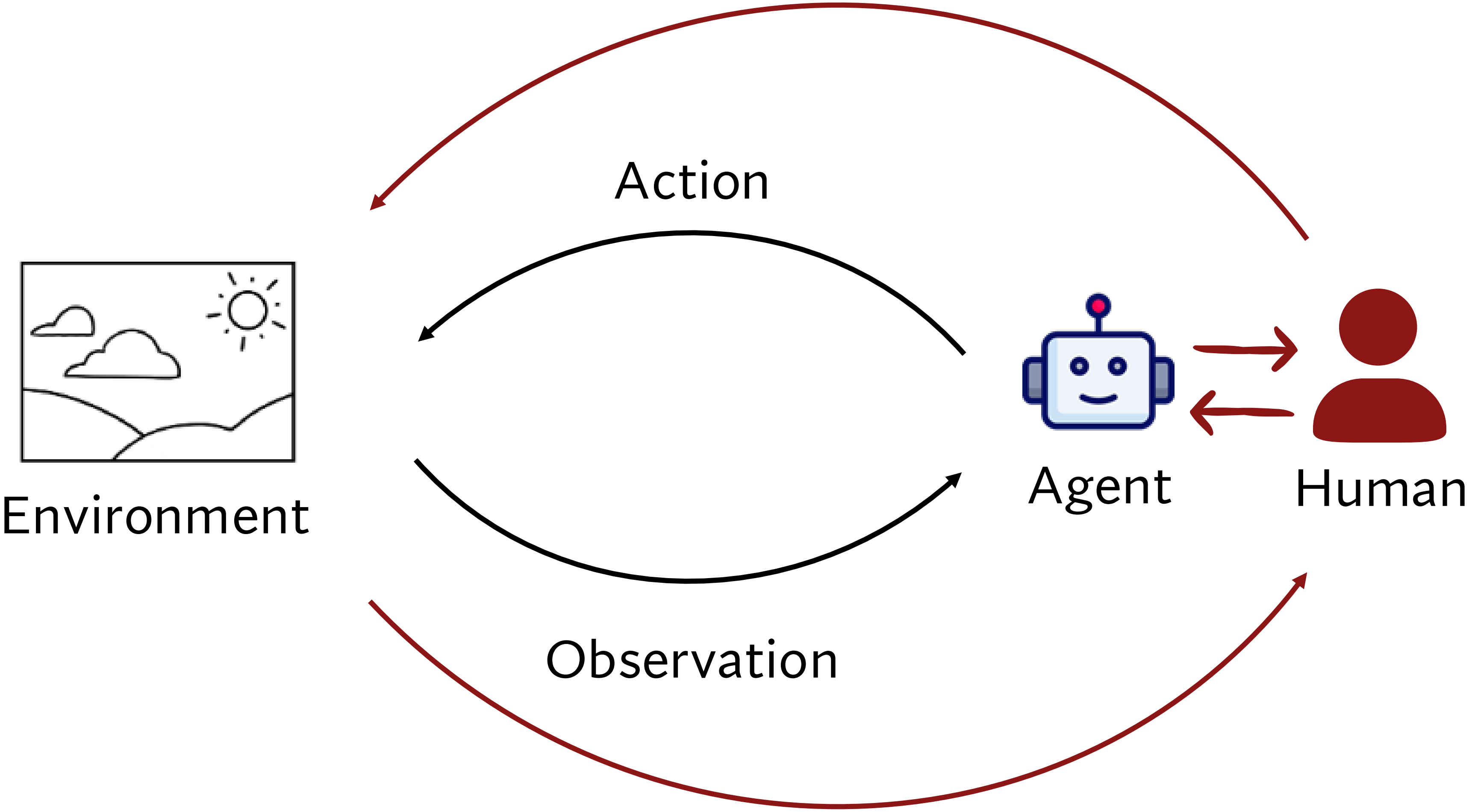
About CollabLLM:

“It helped really well to navigate what to say and what information is needed”

“The AI really helped focusing on one part of the story at a time.”

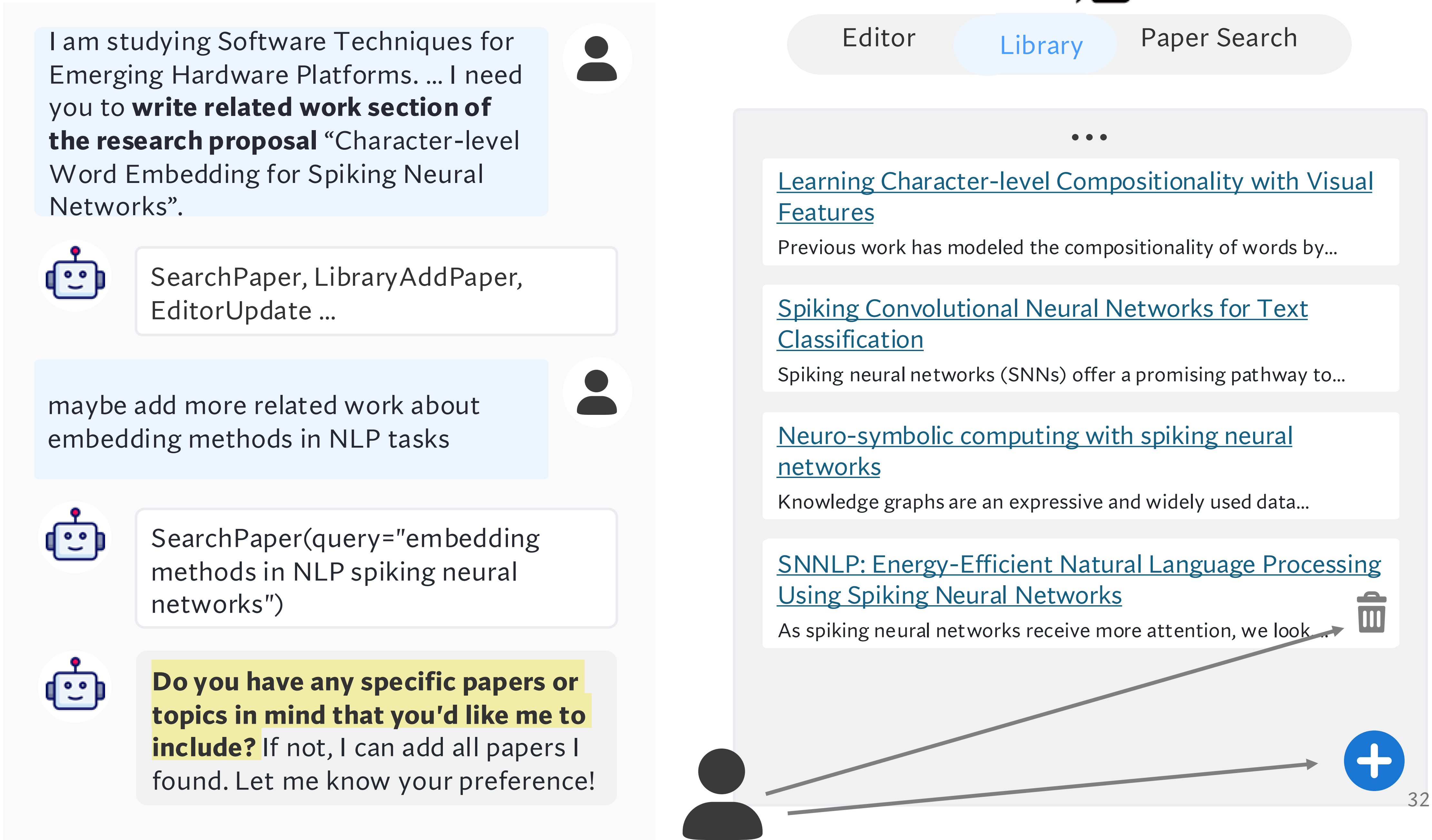
“Asking questions and making you think of things you never thought of”

Enable human-agent collaboration



Shao, Yijia, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. "Collaborative gym: A framework for enabling and evaluating human-agent collaboration." *arXiv preprint arXiv:2412.15701* (2024).

Collaborative Gym



I am studying Software Techniques for Emerging Hardware Platforms. ... I need you to **write related work section of the research proposal** “Character-level Word Embedding for Spiking Neural Networks”.

SearchPaper, LibraryAddPaper, EditorUpdate ...

maybe add more related work about embedding methods in NLP tasks

SearchPaper(query="embedding methods in NLP spiking neural networks")

Do you have any specific papers or topics in mind that you'd like me to include? If not, I can add all papers I found. Let me know your preference!

Learning Character-level Compositionality with Visual Features

Previous work has modeled the compositionality of words by...

Spiking Convolutional Neural Networks for Text Classification

Spiking neural networks (SNNs) offer a promising pathway to...

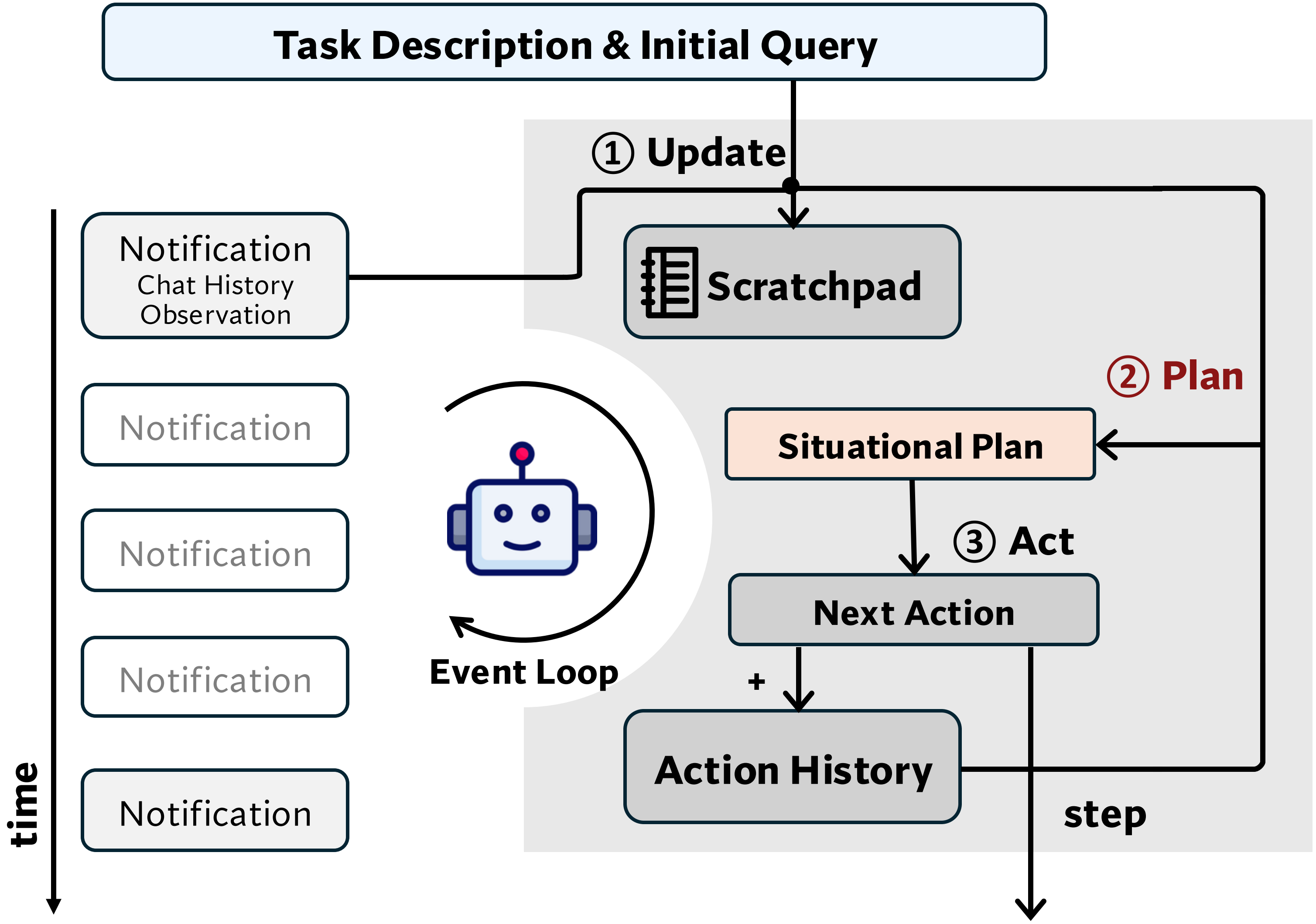
Neuro-symbolic computing with spiking neural networks

Knowledge graphs are an expressive and widely used data...

SNNLP: Energy-Efficient Natural Language Processing Using Spiking Neural Networks

As spiking neural networks receive more attention, we look ...

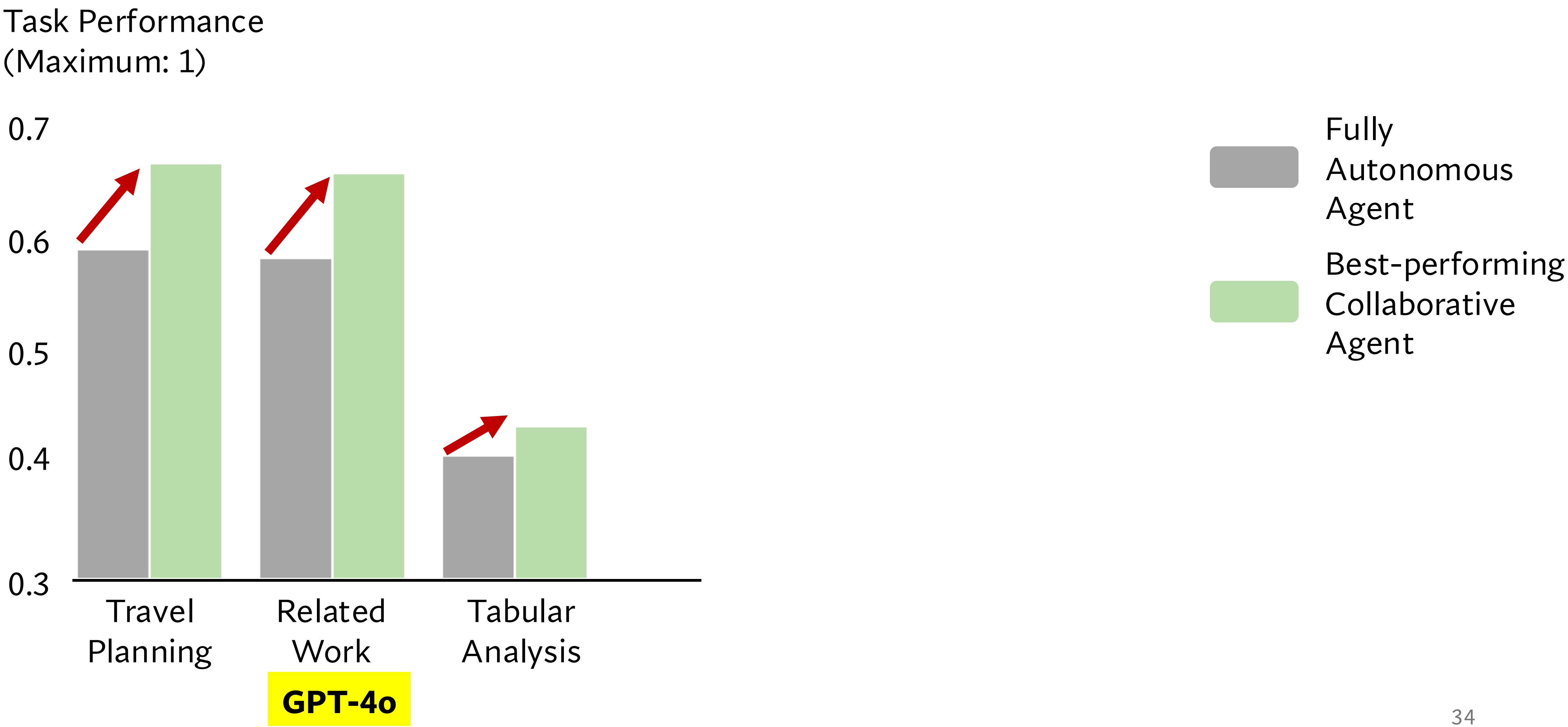
Collaborative Agents w/ Situational Reasoning



Situational Reasoning

- 🤖 Take a task action
- 🤖 Send a message
- 🤖 Defer to human

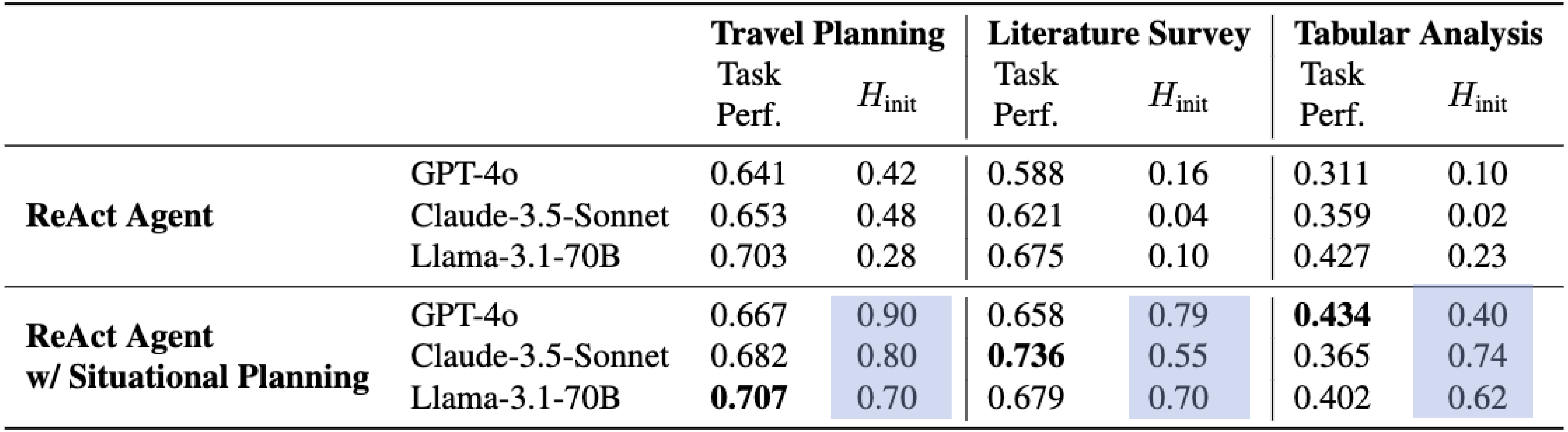
Human-Agent Teams Achieve Better Performances



Audit Human Agent Collaboration Process

Define **initiative entropy** to examine the distribution of initiative taking behaviors

$$H_{\text{init}} = \begin{cases} -\Sigma_{i=1}^{N} p_i \log_N(p_i) & \forall i, p_i > 0, \\ 0 & \exists i, p_i = 0 \end{cases}$$



		Travel Planning		Literature Survey		Tabular Analysis	
		Task Perf.	H_{init}	Task Perf.	H_{init}	Task Perf.	H_{init}
ReAct Agent	GPT-4o	0.641	0.42	0.588	0.16	0.311	0.10
	Claude-3.5-Sonnet	0.653	0.48	0.621	0.04	0.359	0.02
	Llama-3.1-70B	0.703	0.28	0.675	0.10	0.427	0.23
ReAct Agent w/ Situational Planning	GPT-4o	0.667	0.90	0.658	0.79	**0.434**	0.40
	Claude-3.5-Sonnet	0.682	0.80	**0.736**	0.55	0.365	0.74
	Llama-3.1-70B	**0.707**	0.70	0.679	0.70	0.402	0.62

Proactive communication helps human-agent collaboration!

Trajectories Reveal Failures of Today’s Agents

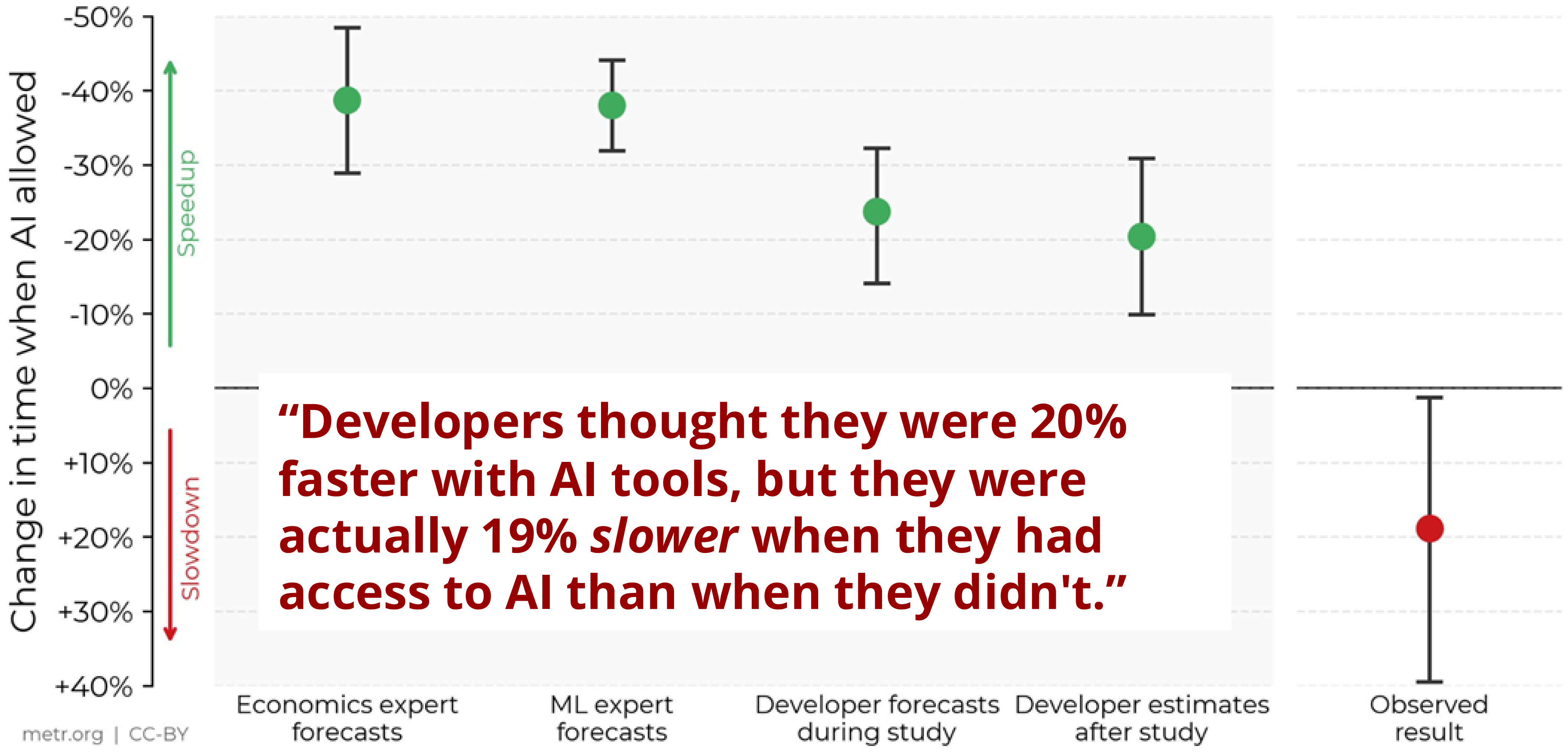
Communication	Agents process tasks **without informing users**	**65%**
Situational Awareness	Agents **disregard session context**, treating each request as an isolated task.	**40%**
Planning	Agents acknowledge tasks but **fail to execute them**.	**39%**
Environment Awareness	Agents **do not assess the feasibility** of requests within constraints of available tools.	**28%**
Personalization	Agents rely on general templates that **do not adapt to individual user needs**	**16%**

Outline

- ✓**What Is Human-AI Interaction** (10 mins)
- ❑**Enable Human-AI Interaction & Collaboration** (20 mins)
  - ✓Automation vs. Augmentation in “Human-AI Collaboration”
  - ✓Agency, RL and situational reasoning to improve collaboration
  - ❑Mixed examples of “does human-AI collaboration work”

Human-AI collaboration “slows down” people

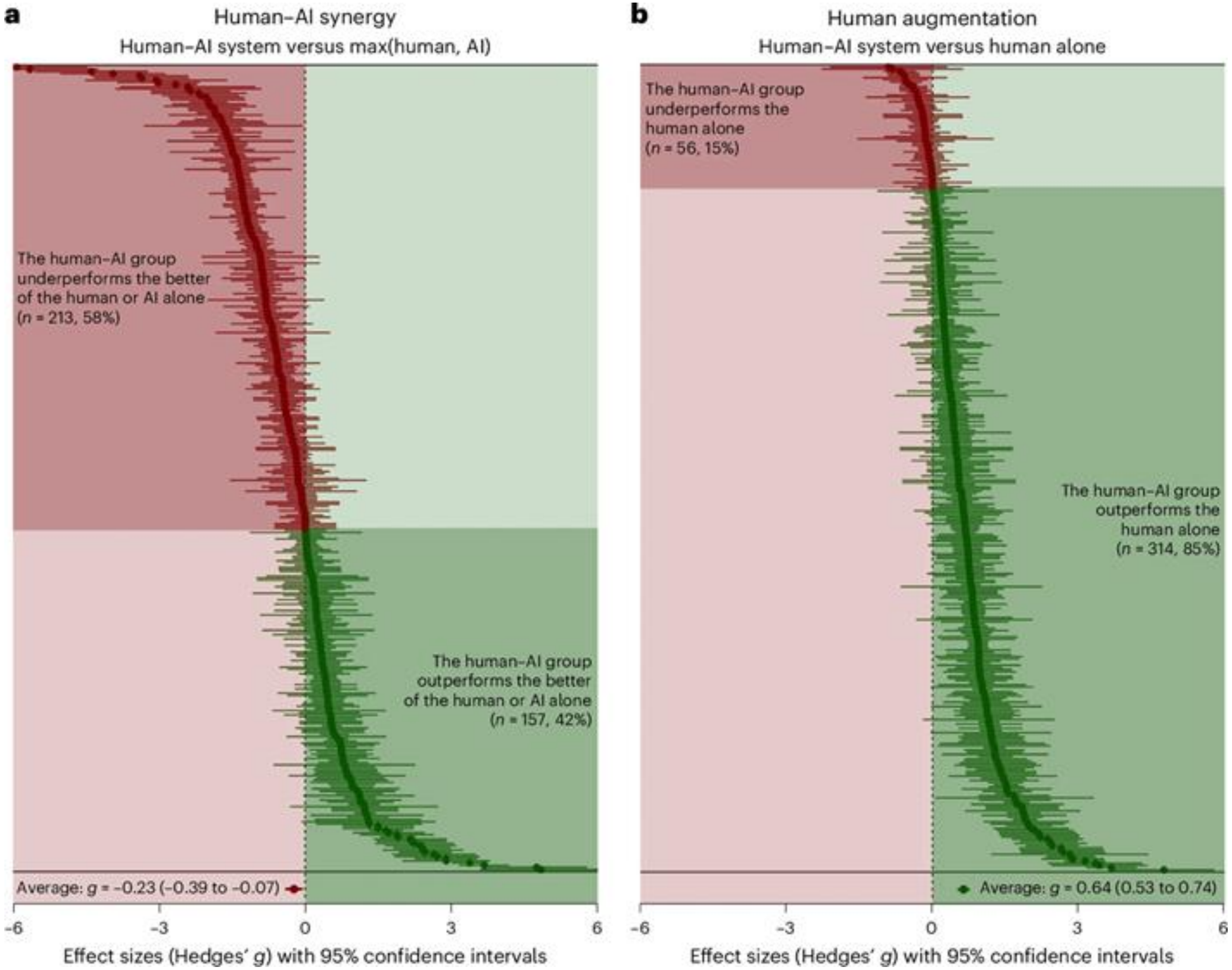
In this RCT, 16 developers with moderate AI experience complete 246 tasks in large and complex projects on which they have an average of 5 years of prior experience.



metr.org | CC-BY

“Developers thought they were 20% faster with AI tools, but they were actually 19% *slower* when they had access to AI than when they didn't.”

When combinations of humans and AI are useful



On average, human–AI combinations performed significantly worse than the best of humans or AI alone

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. "When combinations of humans and AI are useful: A systematic review and meta-analysis." *Nature Human Behaviour* 8, no. 12 (2024): 2293-2303.