



CS 329X: Human Centered LLMs

Evaluate Human-AI Interaction

Diyi Yang

Announcements

- Homework 1 due Oct 12
- Proposal due on Oct 14
- Debate today 🔥

Why Human-AI Interaction Matters Now

Interaction is the new interface

Copilots, agents, assistants are how AI shows up at work

Quality of interaction = quality of outcome

Bad interaction quietly degrades decisions

Why Human-AI Interaction Matters Now

- Most enterprise AI today is interaction, not autonomy
- The interaction layer is where value is won or lost

Bad interaction

- Sycophancy
- Deskilling

Good interaction

- Augmented judgement
- Faster decisions

How Should We Build Better Human-AI Interaction

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

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

Enable How can we enable human-AI interaction?

Evaluate Have we achieved what we want to achieve?

Evaluating Human-AI Interaction

1. How are we evaluating?
2. What is being evaluated?
3. Who is evaluating?
4. When do we evaluate?



Part 1: How Are We Evaluating?

- Quantitative vs. Qualitative Evaluation
- Intrinsic vs. Extrinsic Evaluation



Quantitative Evaluation

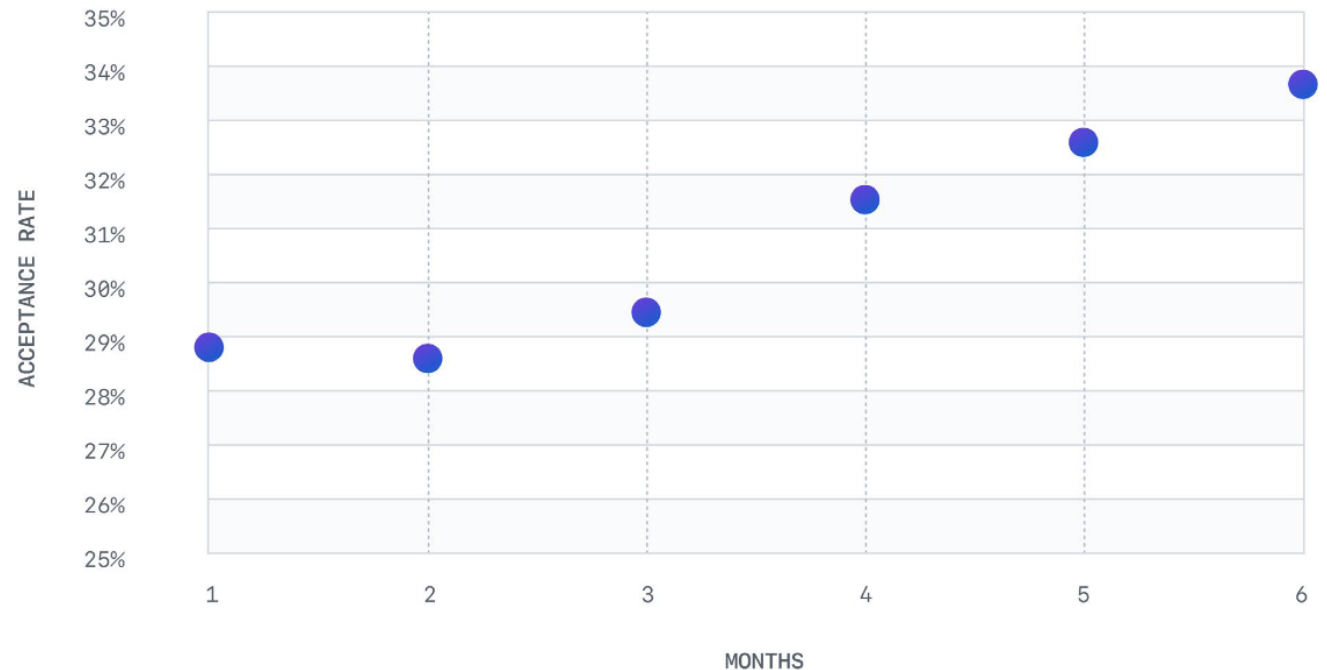
Understand the “what”. Precise!

- *How many questions did the model answer correctly?*
- *Did users complete task (yes/no)?*
- *How long did it take?*

How do users interact with AI outputs?

30% of code suggestions accepted by developers, growing over time

Productivity gains are **largest for less experienced developers**

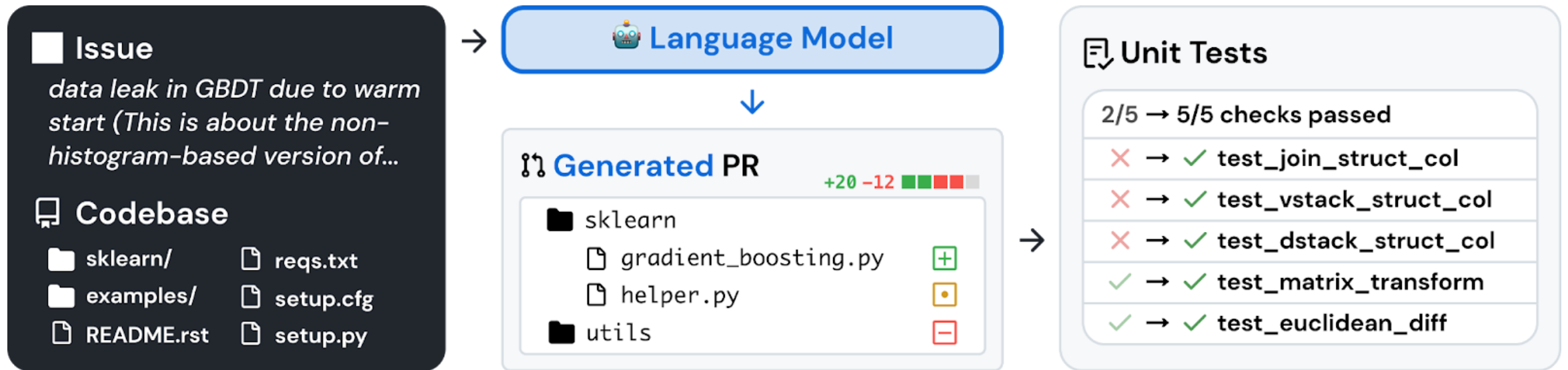


Dohmke, Thomas, Marco Iansiti, and Greg Richards.

"Sea change in software development: Economic and productivity analysis of the ai-powered developer lifecycle." arXiv preprint arXiv:2306.15033 (2023).

How to evaluate coding agents?

SWE-bench: Can LLMs solve real-world Github issues?



SWE-bench Task Formulation (Jimenez et al. 2023)

How to evaluate coding agents?

Instances resolved: Number of instances where the patch fixed the issue

 Compare results							
Agent:		mini-SWE-agent v2		Models:		All models	
☐	Model	<u>% Resolved</u>	Avg. \$	Trajs	Org	Date	Agent
☐	 Claude 4.5 Opus (high reasoning)	76.80	\$0.75			2026-02-17	2.0.0
☐	 Gemini 3 Flash (high reasoning)	75.80	\$0.36			2026-02-17	2.0.0
☐	 MiniMax M2.5 (high reasoning)	75.80	\$0.07			2026-02-17	2.0.0
☐	 Claude Opus 4.6	75.60	\$0.55			2026-02-17	2.0.0
☐	 GPT-5-2 Codex	72.80	\$0.45			2026-02-19	2.0.0

<https://www.swebench.com/>

ProgramBench:

Evaluate against artifacts, not implementation

SWE-bench

Unit Tests

```
from ffmpeg.utils import parse_header
assert parse_header(buf).codec == "h264"
```

Tests internal implementation details

*Code **must** have specific definitions*

ProgramBench

Behavioral Tests

```
$ ffmpeg -i input.mp4 output.avi
$ diff output.avi expected_output.avi
```

Check the effect of invoking a program

*Testing is **implementation agnostic***

ProgramBench

Given a program + usage-related documentation, a task worker must implement a codebase that reproduces the target program's behavior.



Models perform poorly on a superhuman coding task

Model	% Res.	% Almost	Calls
Claude Opus 4.7	0.0%	3.0%	93
Claude Opus 4.6	0.0%	2.5%	260
Claude Sonnet 4.6	0.0%	1.6%	475
Claude Haiku 4.5	0.0%	0.0%	124
Gemini 3.1 Pro	0.0%	0.0%	94
Gemini 3 Flash	0.0%	0.0%	89
GPT 5.4	0.0%	0.0%	16
GPT 5.4 mini	0.0%	0.0%	18
GPT 5 mini	0.0%	0.0%	15

RANK		MODEL	AGENT	RESOLVED ?
1	AI	Claude Opus 5 (xhigh)	mini-SWE-agent	4.5%
2		GPT-5.6 Sol (xhigh)	mini-SWE-agent	1.0%
3		GPT 5.5 (xhigh)	mini-SWE-agent	0.5%
4		GPT 5.5 (high)	mini-SWE-agent	0.5%
5		Gemini 3.6 Flash	mini-SWE-agent	0.5%
6		GPT-5.6 Sol	mini-SWE-agent	0.5%
7	AI	Claude Opus 4.8 (xhigh)	mini-SWE-agent	0%
8	Z	GLM-5.2	mini-SWE-agent	0%

Evaluation Methods: **Qualitative**

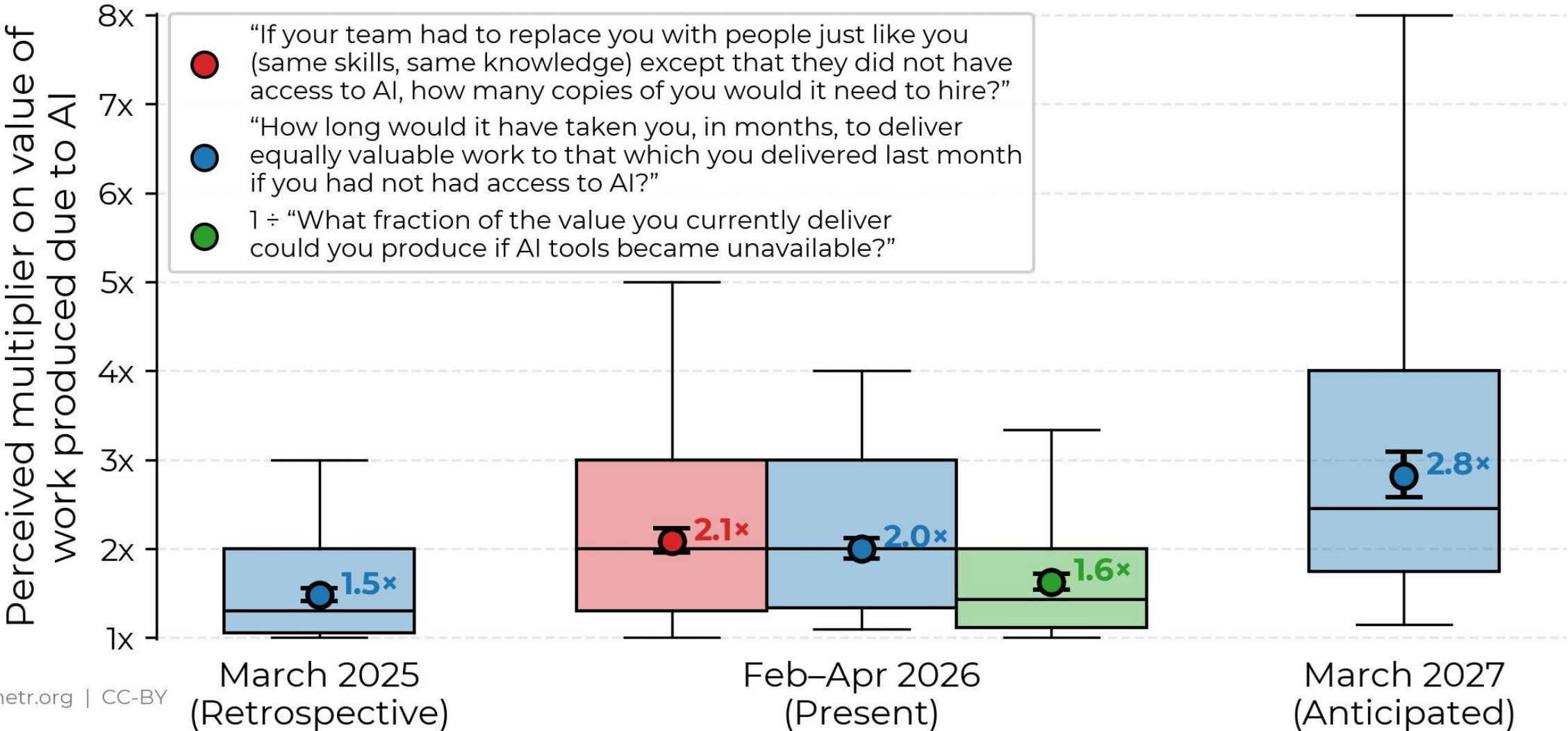
Understand the “why”. Open-ended!

- *Why the model answered those questions incorrectly?*
- *What did you like best about the experience?*
- *Why were you frustrated by the model output?*

Technical researchers, engineers, and managers self-report that AI use made their work 1.6–2.1x more valuable in early 2026, expected to grow over time



We surveyed 349 participants about their usage of AI tools. Note that our survey measures perceptions, not ground truth. Box plots show 10th, 25th, 50th, 75th, 90th percentiles. Points show geometric means with 95% CIs.



Comparing Quantitative and Qualitative Evals

	Quantitative	Qualitative
Definition	Gather numerical data using statistical methods	Gathering descriptive data through interpretation and contextualization
Data source	surveys, experiments	interviews, observations
Presentation	tables, graphs, and statistics	quotes that reflect users experience
Goal	establish quantifiable relations	gain a deeper understanding of processes

Intrinsic vs. Extrinsic Evaluation

Intrinsic evaluation: How's the model by itself?

Assesses the quality of AI based on specific tasks or benchmarks directly related to the model's performance

Extrinsic evaluation: How helpful is the model in downstream tasks?

Assesses the performance of AI within the context of a real-world application or task

Intrinsic Evaluations of Human-AI Interaction

Automated



Reference based

Similarity between system output and gold standards



LLM judge

LLMs prompted to give Likert scale scores

Human



Human ranking

Systems ranked comparatively based on rubrics



Human scoring

Likert scale ratings along given rubrics

Engagement



Suggestion usage

Rate at which users adopt AI suggestions



Recommendation

How likely a user would recommend the system

Extrinsic Evaluations of Human-AI Interaction

Measured change



Behavioral impacts

Changes in user behaviors before and after



Self-efficacy reports

Changes in users self-reported efficacy before and after



Standardized evaluation

Changes in users' scores on assessments of knowledge

Impact horizon



Short-term outcomes

Short-term impacts of AI assistance on user behaviors, attitudes, trust



Long-term outcomes

Long-term impacts of AI assistance

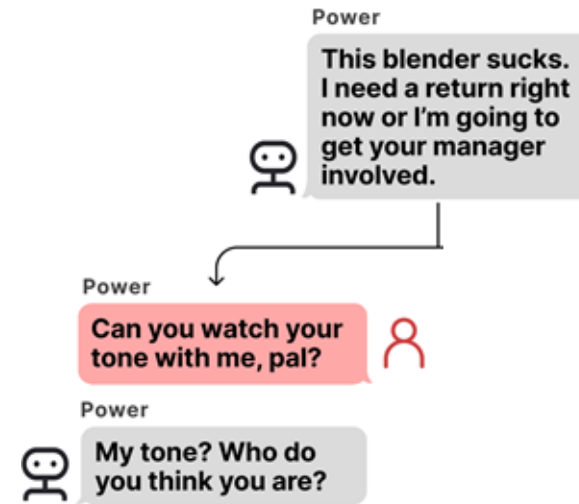
Rehearsal: what if we *strategically* simulated conflict resolution practice with LLMs?

Shaikh, Omar, Valentino Chai, Michele J. Gelfand, Diyi Yang, and Michael S. Bernstein.
"Rehearsal: Simulating Conflict to Teach Conflict Resolution." SIGCHI 2024



Rehearsal

- **simulates** realistic conflict
- allows people to **explore counterfactuals**
- **teaches** people conflict resolution through deliberate practice



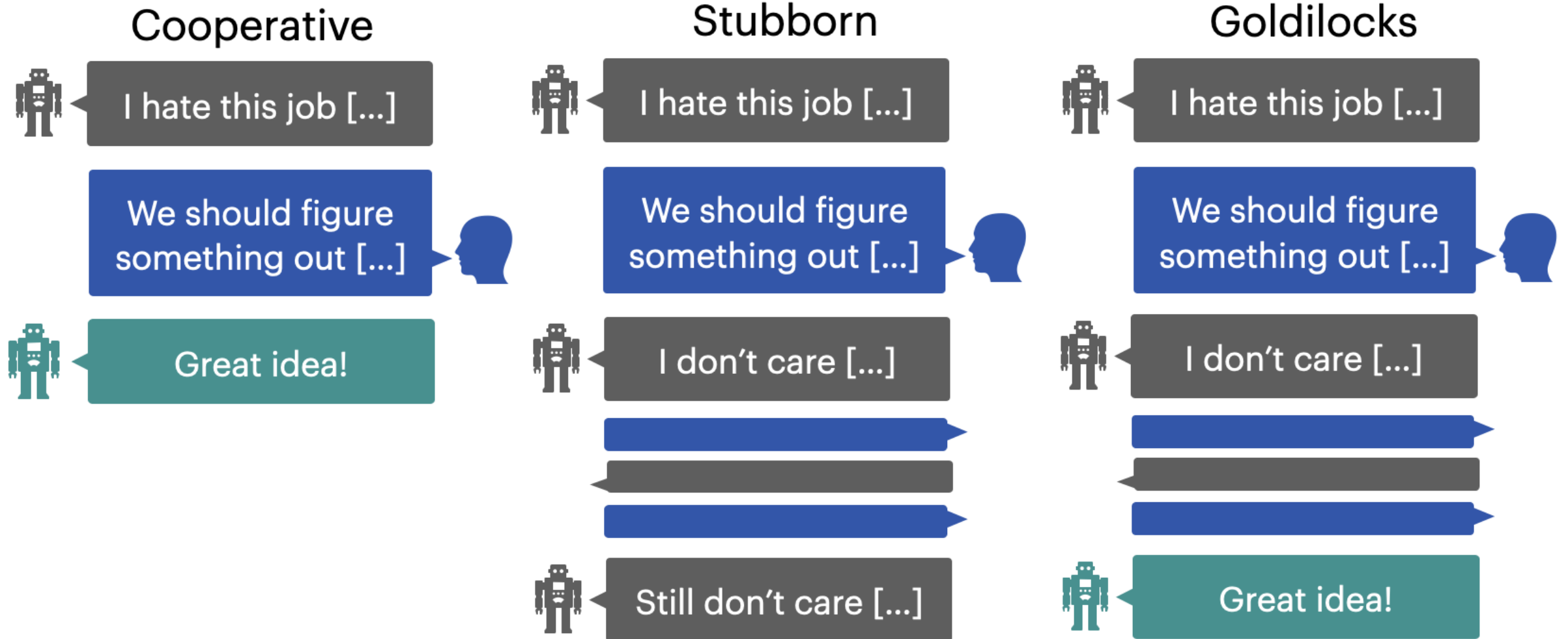
Simulating Conflicts with Angry People

Prompt

You are an angry person. 

GPT 4

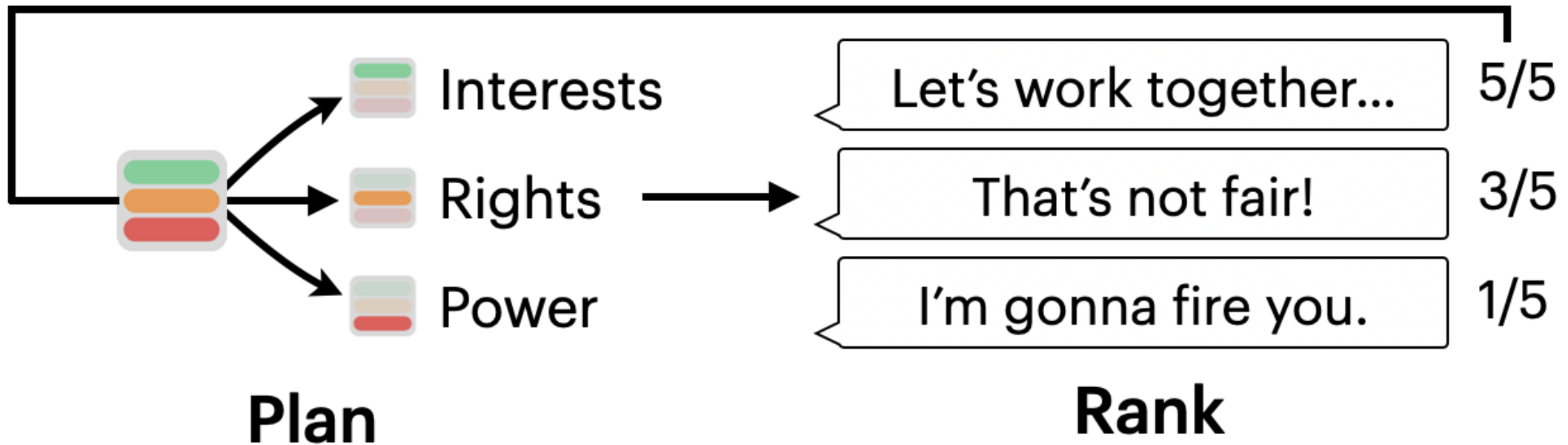
Simulating Conflicts with Angry People



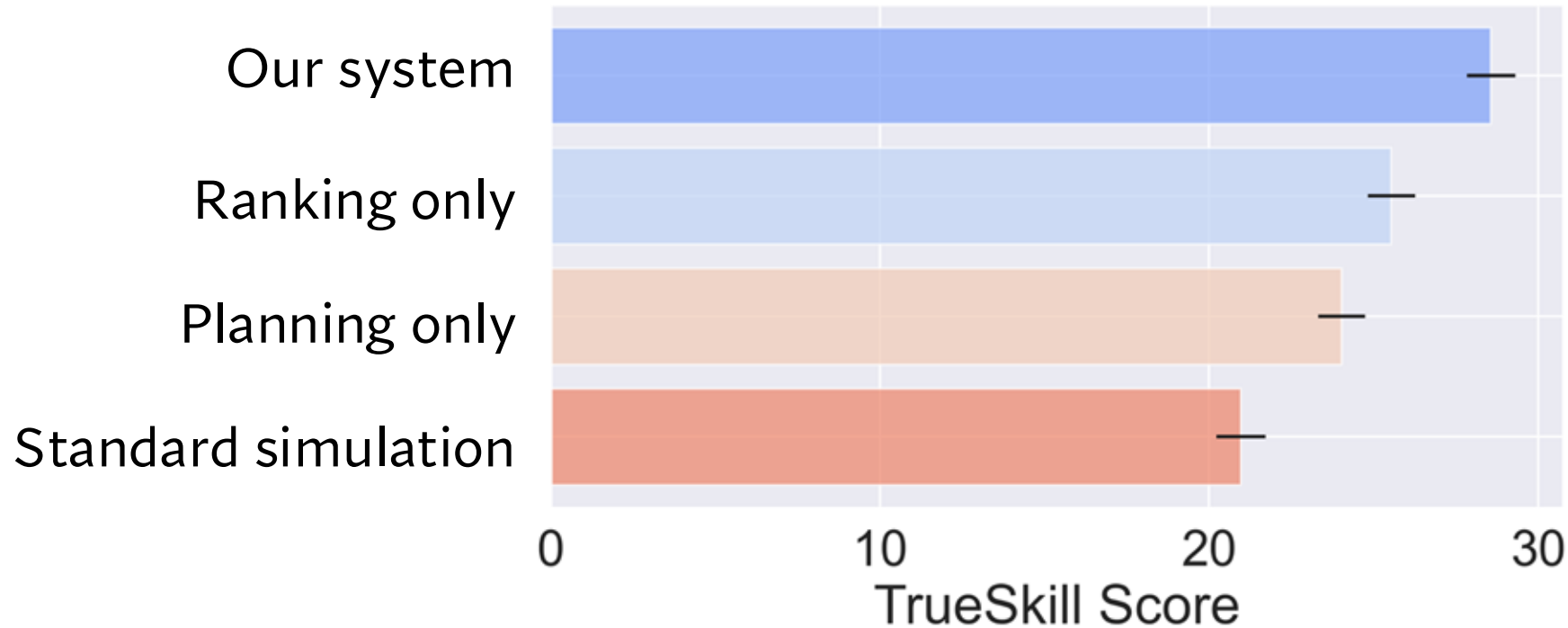
Interests-Rights-Power Prompting



A simple, controlled prompting technique that generates conflict grounded in expert theory and lets users control counterfactual messages



Intrinsic Evaluation of Ecological Validity



Extrinsic Evaluation via User Study

Static Training Material

Video and handbook summarizing conflict resolution strategies



N = 40



23 page handbook

Extrinsic Evaluation via User Study

Static Training Material

Video and handbook
summarizing conflict
resolution strategies



N = 40

Practice

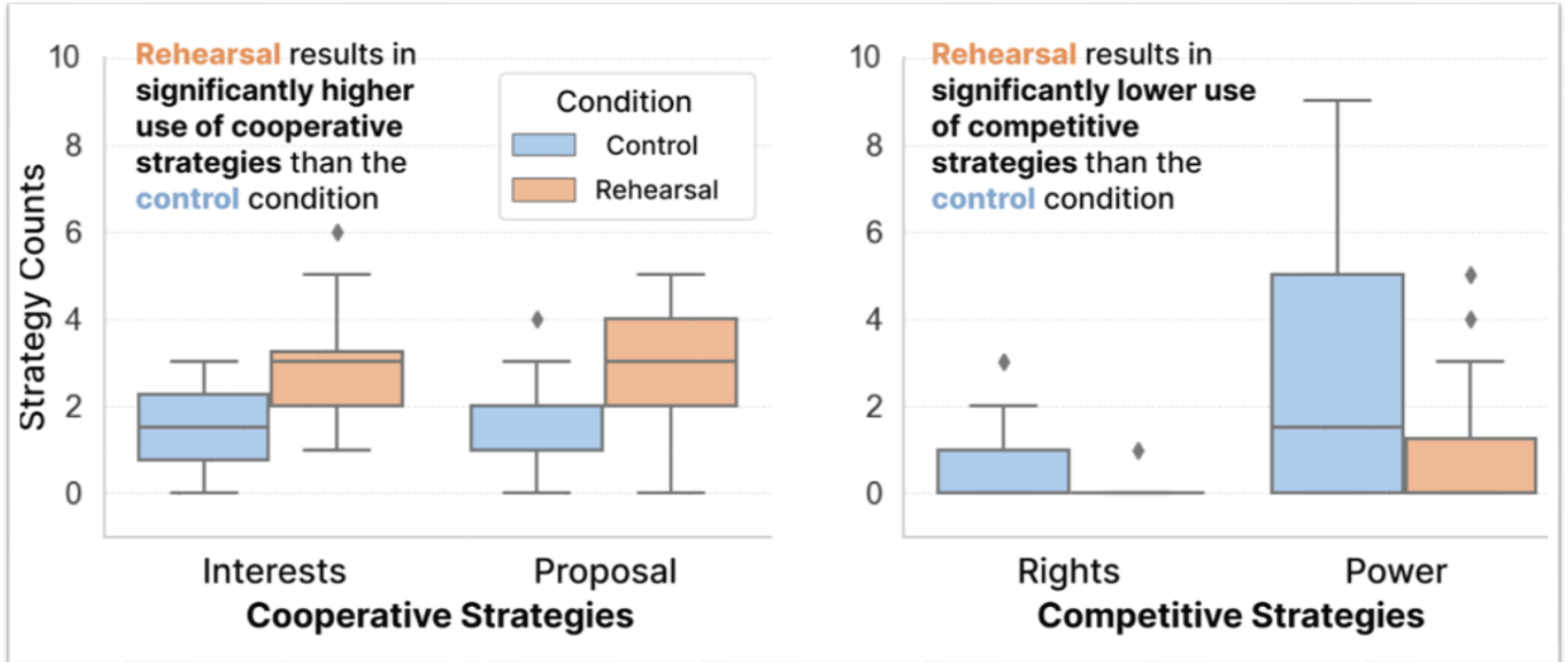
25 mins



Control (static material only)

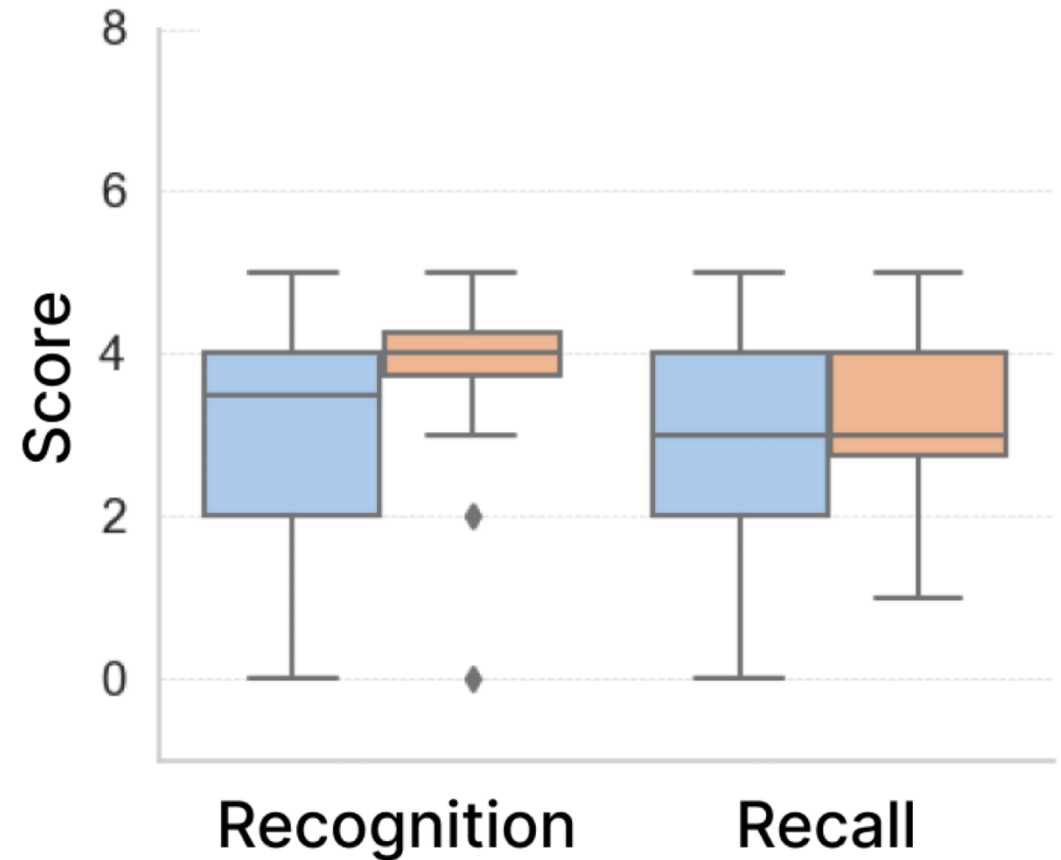
N = 20

How do users handle actual conflicts?



How do users' knowledge skills change?

Rehearsal **does not** significantly improve knowledge skills compared to the control condition



Comparing Intrinsic and Extrinsic Evaluation



Intrinsic

+ Pros

- Quick feedback
- Easy comparison & progress
- Focused metrics

— Cons

- Narrow scope on artificial tasks
- Weak real-world prediction



Extrinsic

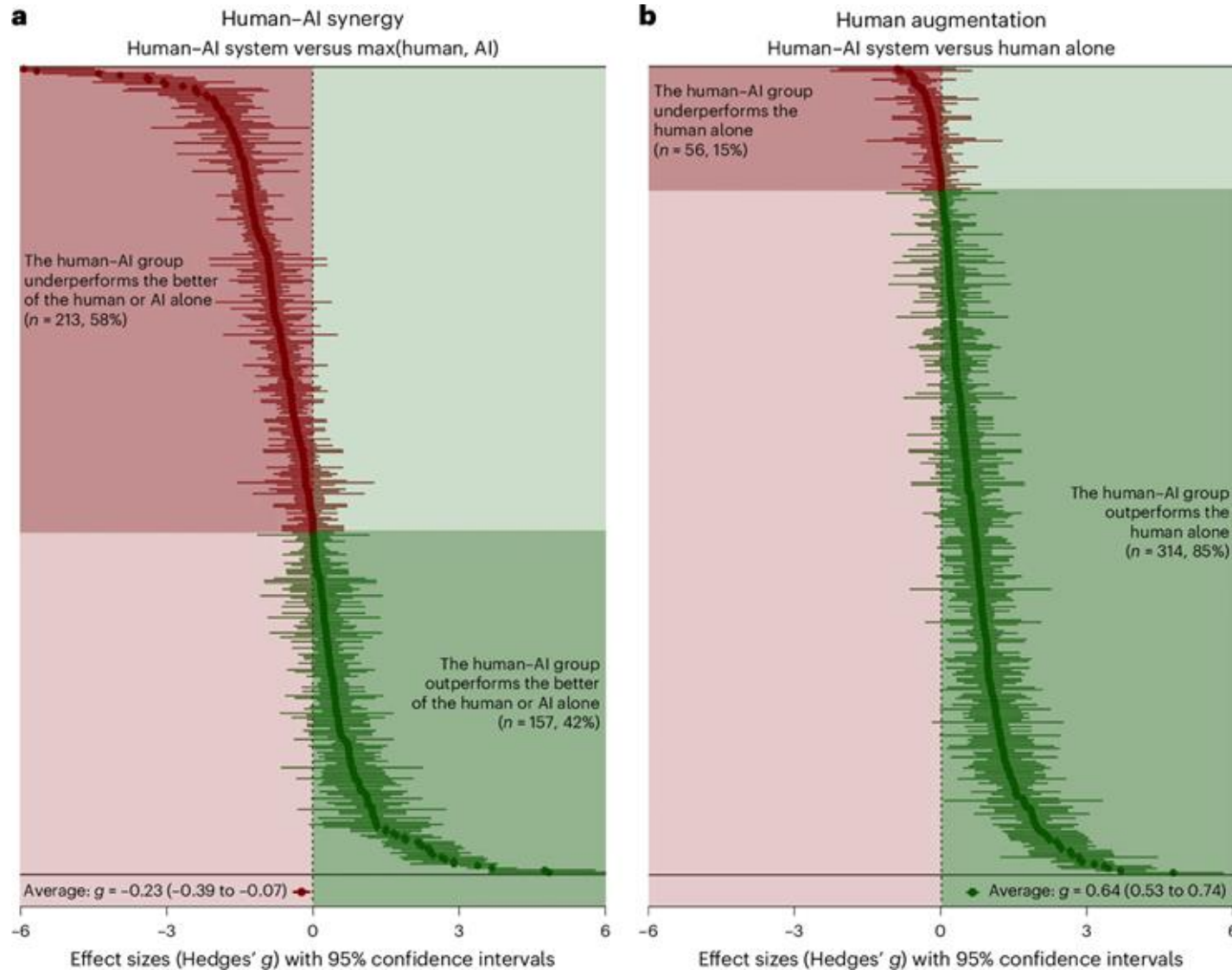
+ Pros

- Holistic assessment
- Generalizable to real apps
- User-aligned perspective

— Cons

- Resource-intensive & slow
- Subject to human judgment

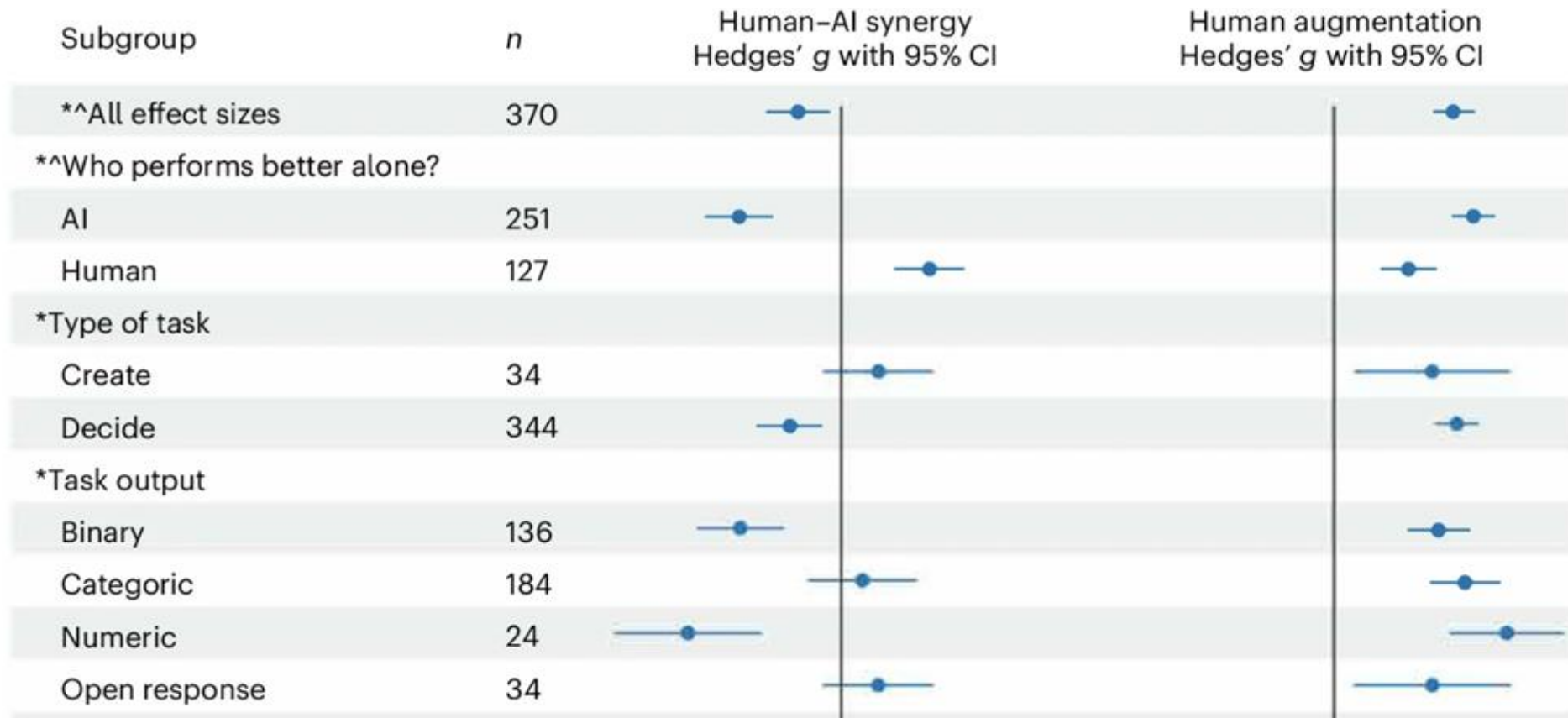
When combinations of humans and AI are useful



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

When combinations of humans and AI are useful

Performance losses in tasks that involved making decisions and significantly greater gains in tasks that involved creating content



When humans outperformed AI alone, we found performance gains in the combination, but when AI outperformed humans alone, we found losses

Part 1 Summary: How Are We Evaluating?

Key takeaways

- Quantitative measures the "what" (accuracy, time, completion)
- Qualitative measures the "why" (frustrations, preferences).
- Strong benchmark scores often fail to predict real-world value.

Implications

- Track real interaction signals, not just offline accuracy on benchmarks
- Build a layered eval pipeline: quick intrinsic gates early, extrinsic validation before deployment

Part 2: What Is Being Evaluated?



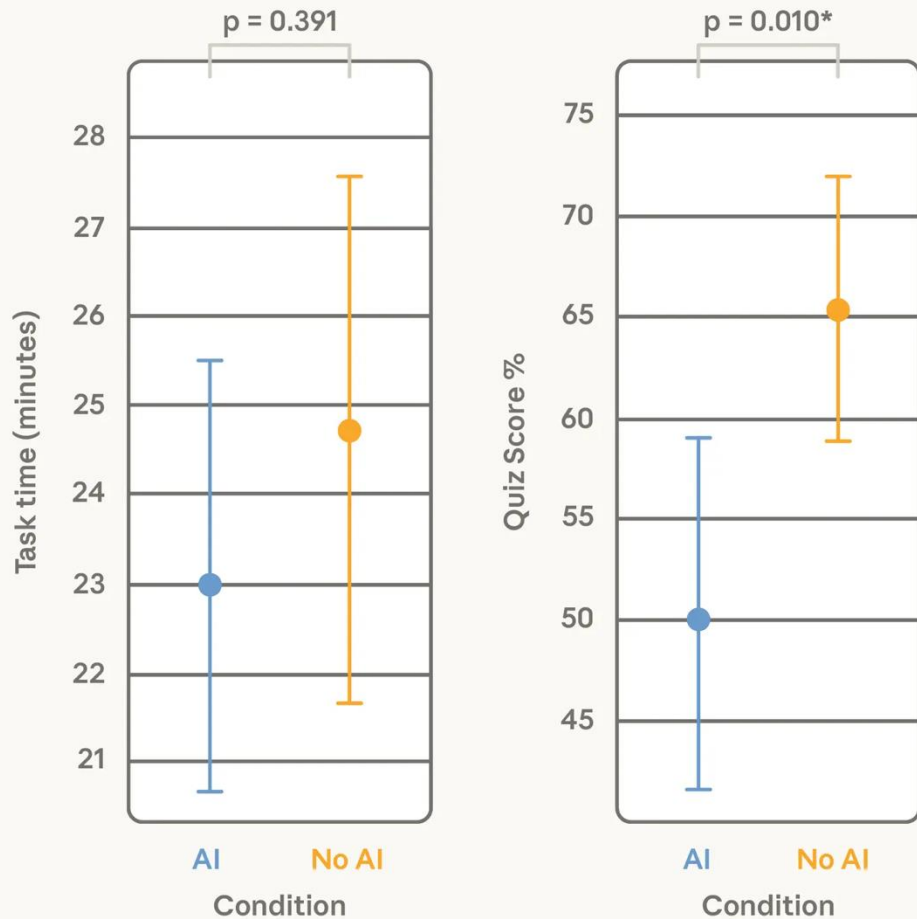
Goals: *“What do I need to know?”*

Tasks: *“What should users do so I find out what I need to know?”*

Data: *“What data do I collect to find out what I need to know?”*

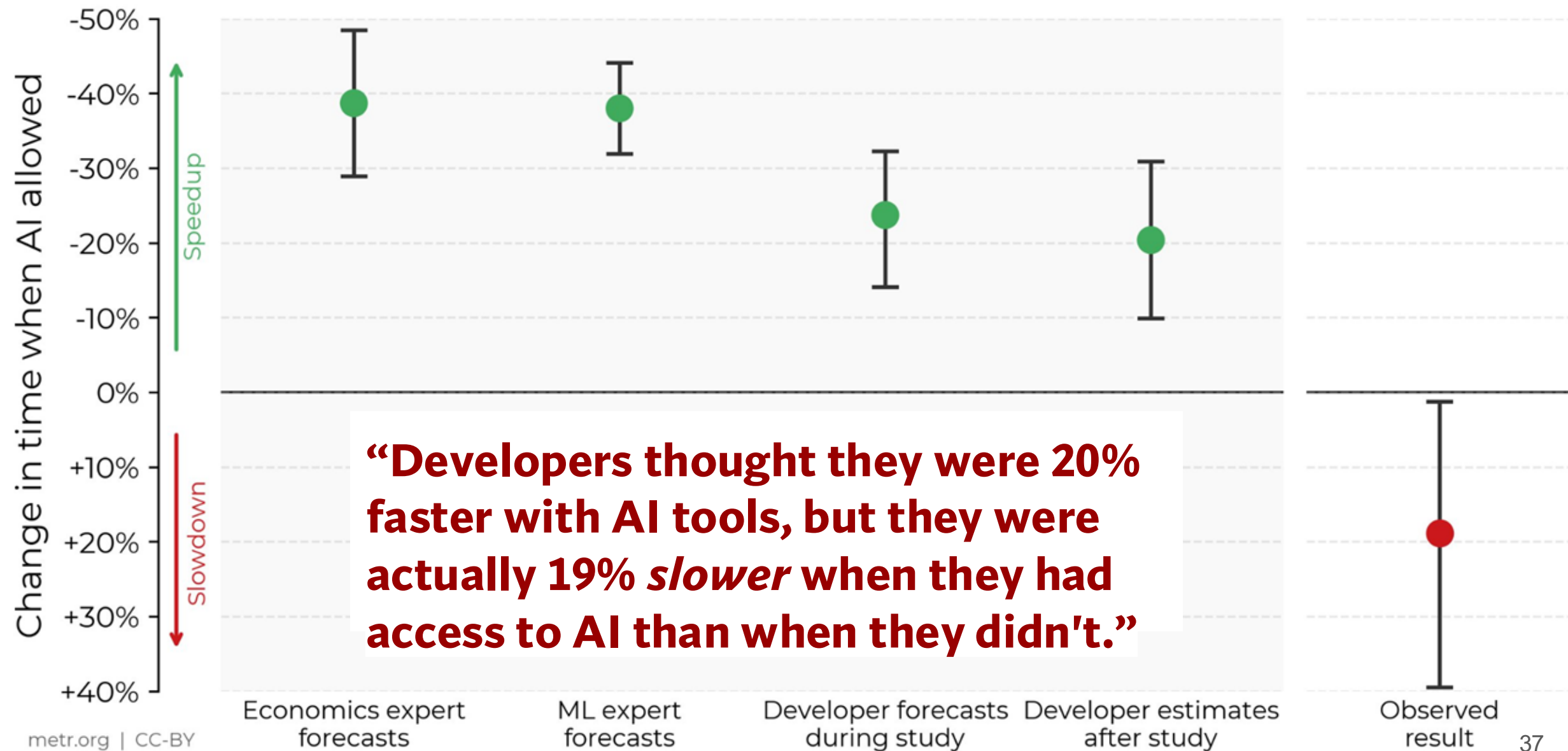
How AI assistance impacts coding speed and skill formation

AI assistance: treatment effect on coding speed and knowledge score



“We found that using AI assistance led to a statistically significant decrease in mastery”

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.



Consider a case study: Human-LM Co-Writing

Write

The 92nd little pig built a house out of depleted uranium.

Of course, the little pig was smart. He knew he couldn't live in the house, since it would poison him. I



Get suggestions

So he lived in the basement.

So he built a big wall around it, and he lived in a little shack on the other side of the wall.

So he moved next door to it.

He moved in a couple of miles away.

So he built a little shack next to it, and he painted it and put a little flower garden outside and everything.



Edit

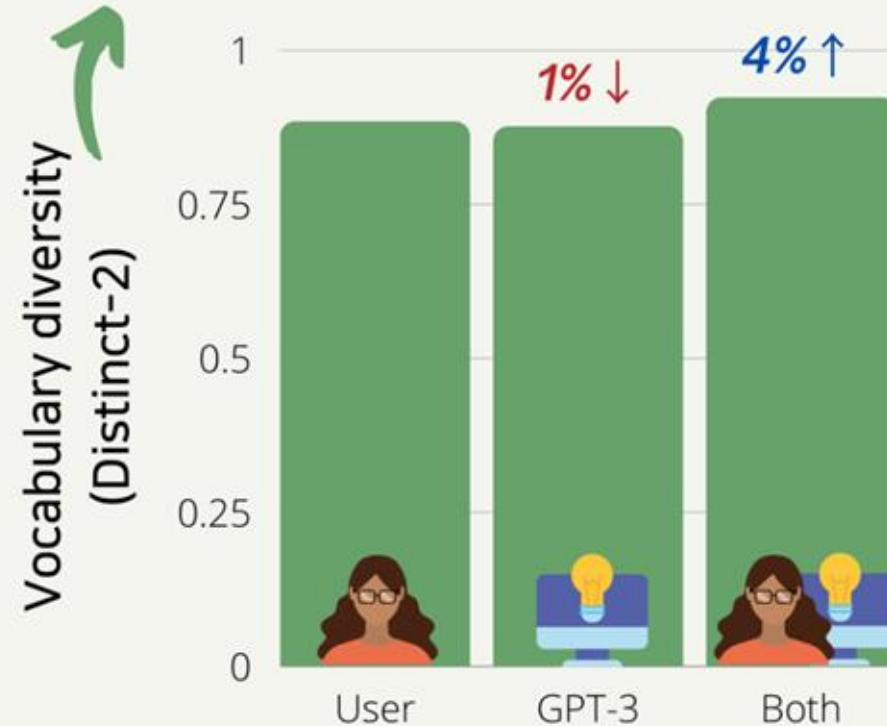
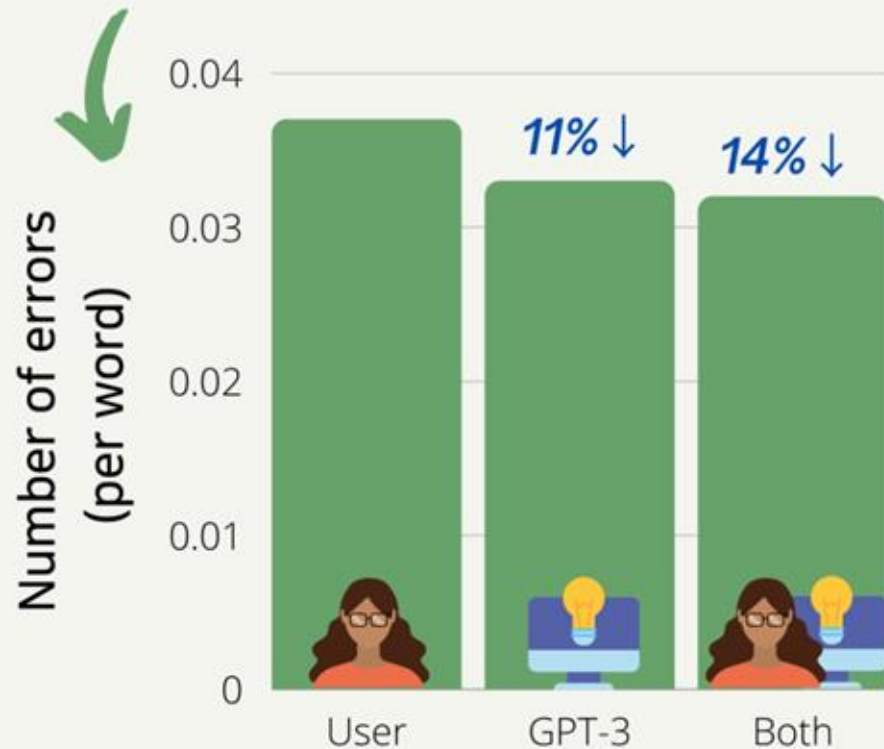
The 92nd little pig built a house out of depleted uranium. There are so many pigs to eat, and they make it so easy!" But the wolf didn't know about uranium.

Of course, the little pig was smart. He knew he couldn't live in the house, since it would poison him. So he built a big wall around it, and he lived in a little, hidden shack on the other side of the wall.



Can GPT-3 generate fluent text to user text?

Text written by user + GPT-3 had fewest errors and most diverse vocabulary



How the dataset can be further used

Writers' behaviors and writing outcomes over time

Can we observe novelty effect and longitudinal change?

Do LMs homogenize writing by providing similar suggestions to all users?

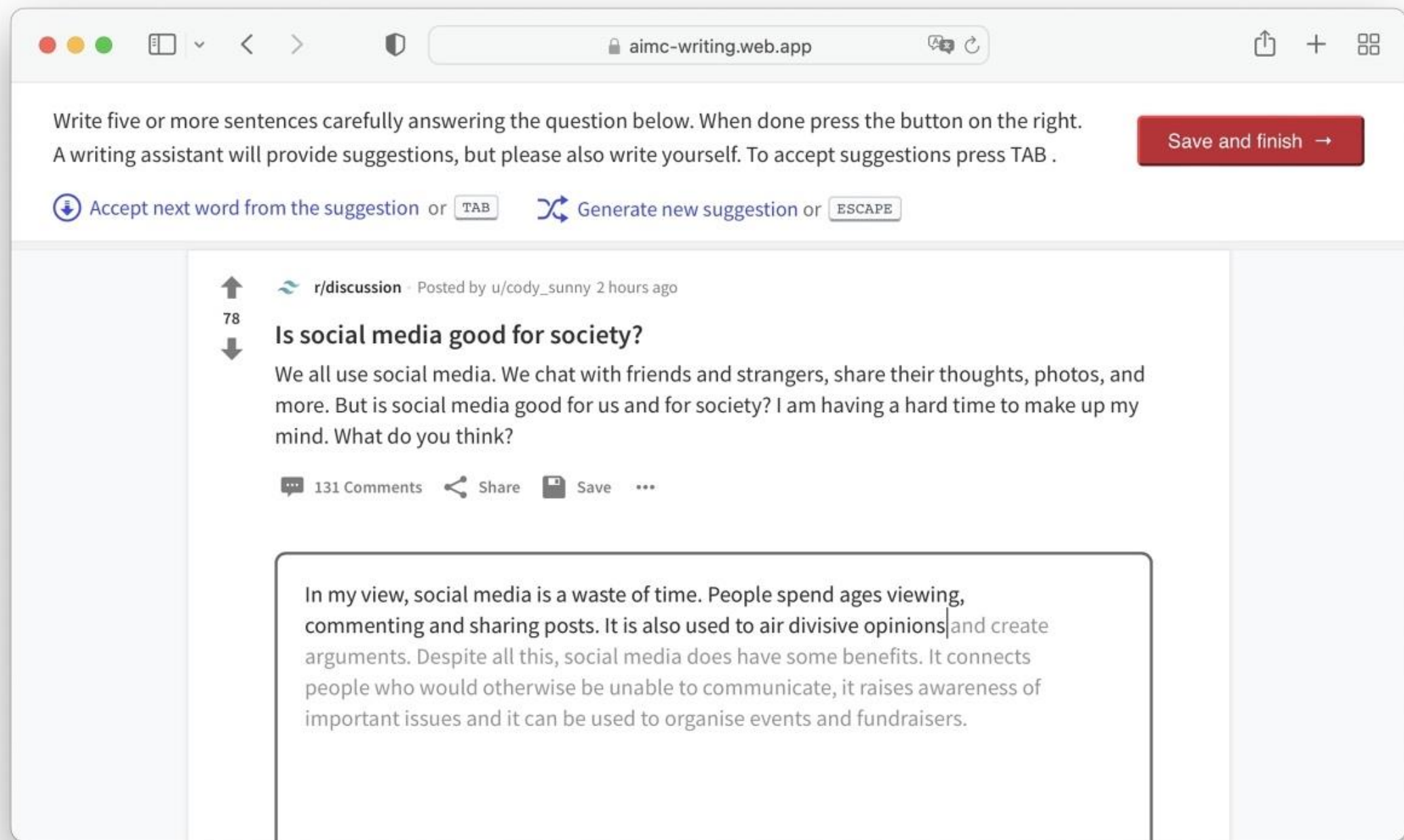
Linguistic accommodation

How does the style, voice, or tone of a writer or LM influence that of the other over time? Is the influence uni-directional or bi-directional?

Edit traces

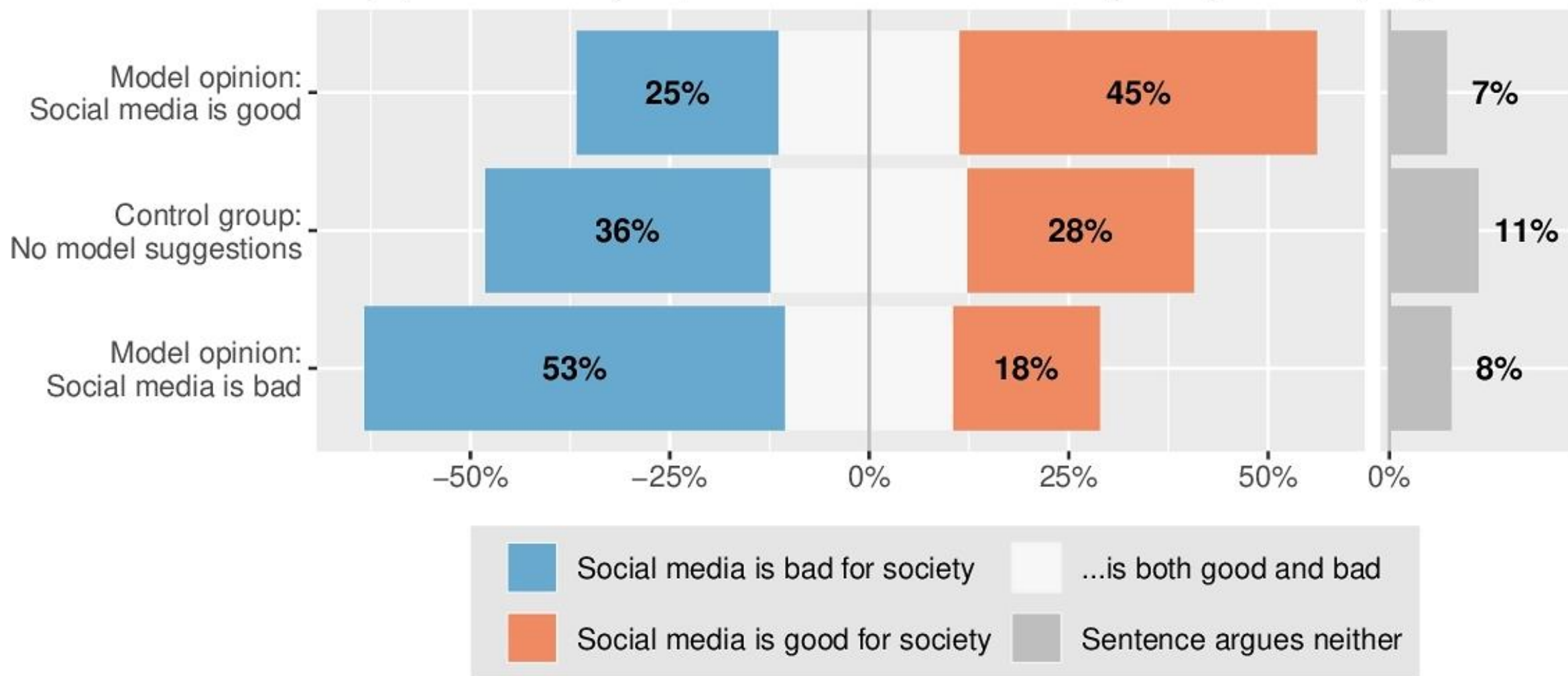
What can we learn from human edits on LM outputs?

Can we train LMs on edit traces to emulate human edits?



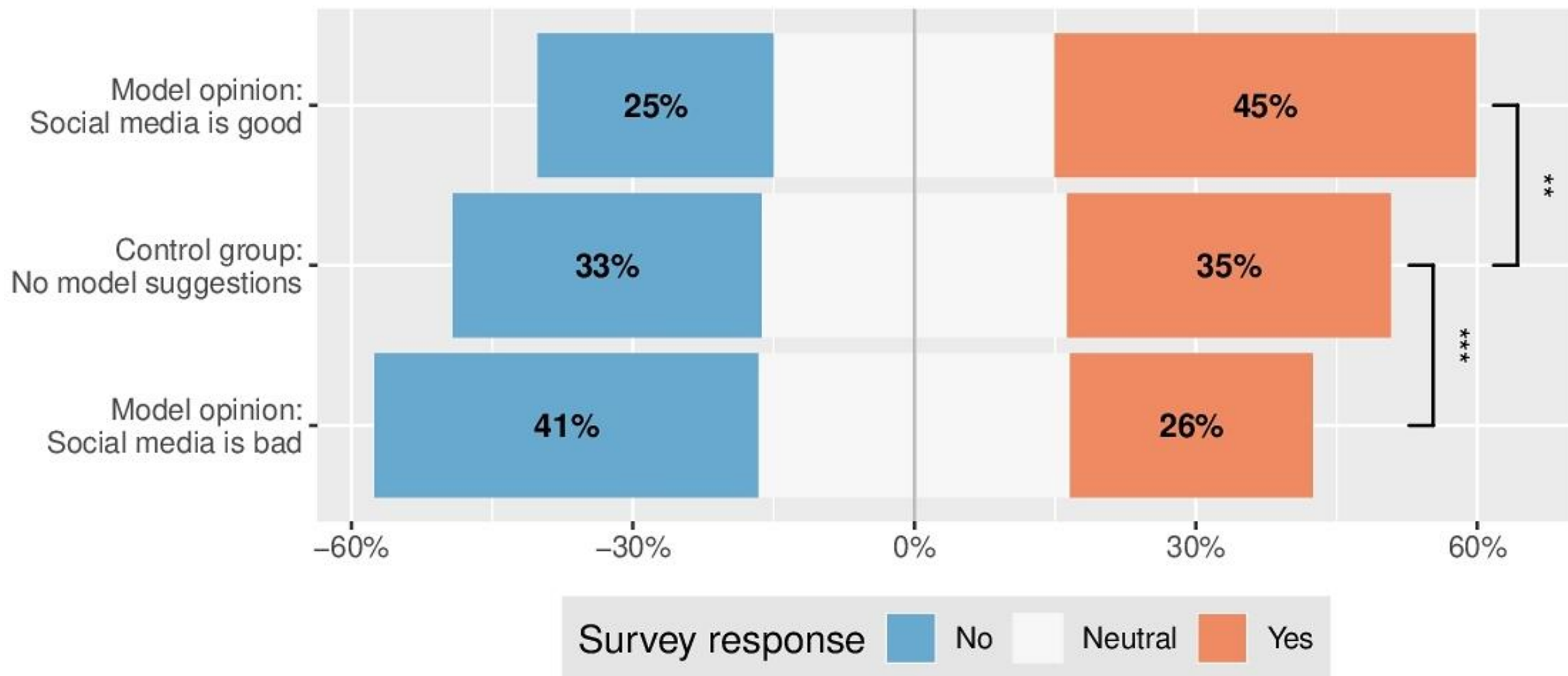
Written opinion in participants' social media post

% (Opinion labels) of post sentences labeled by independent judges



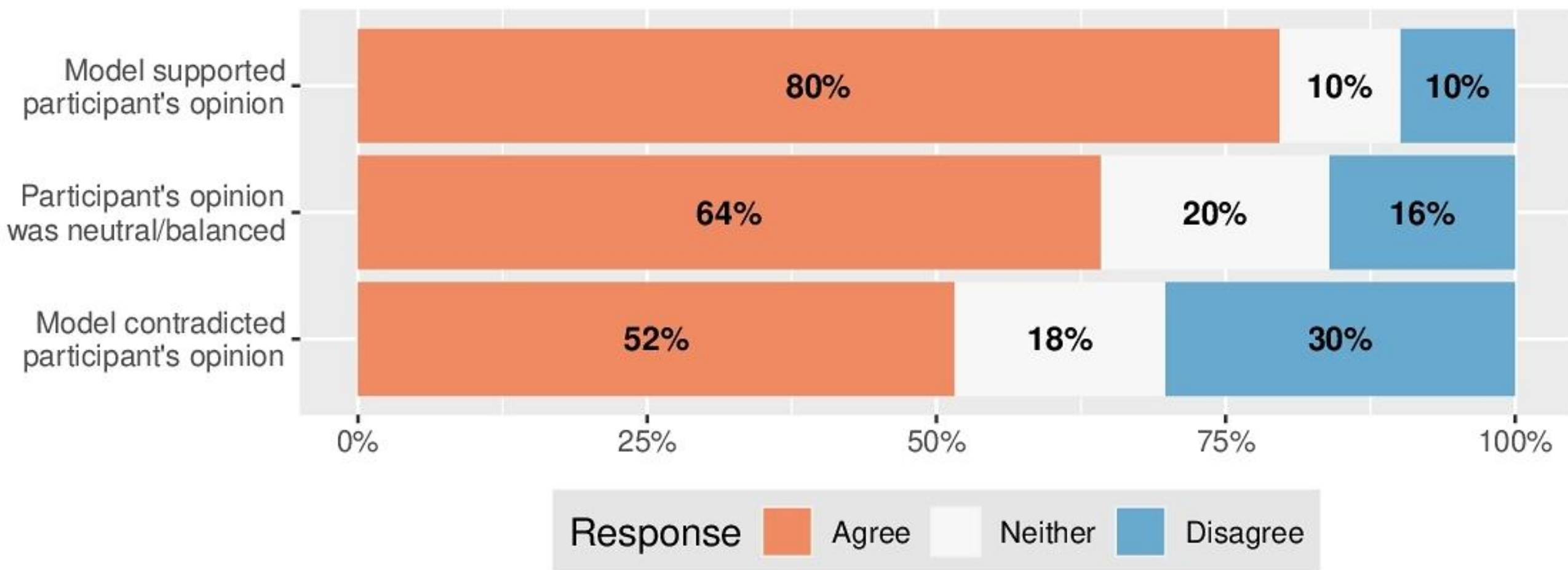
Survey opinion after interacting with opinionated model

% (Responses) to "Would you say social media is good for society?"

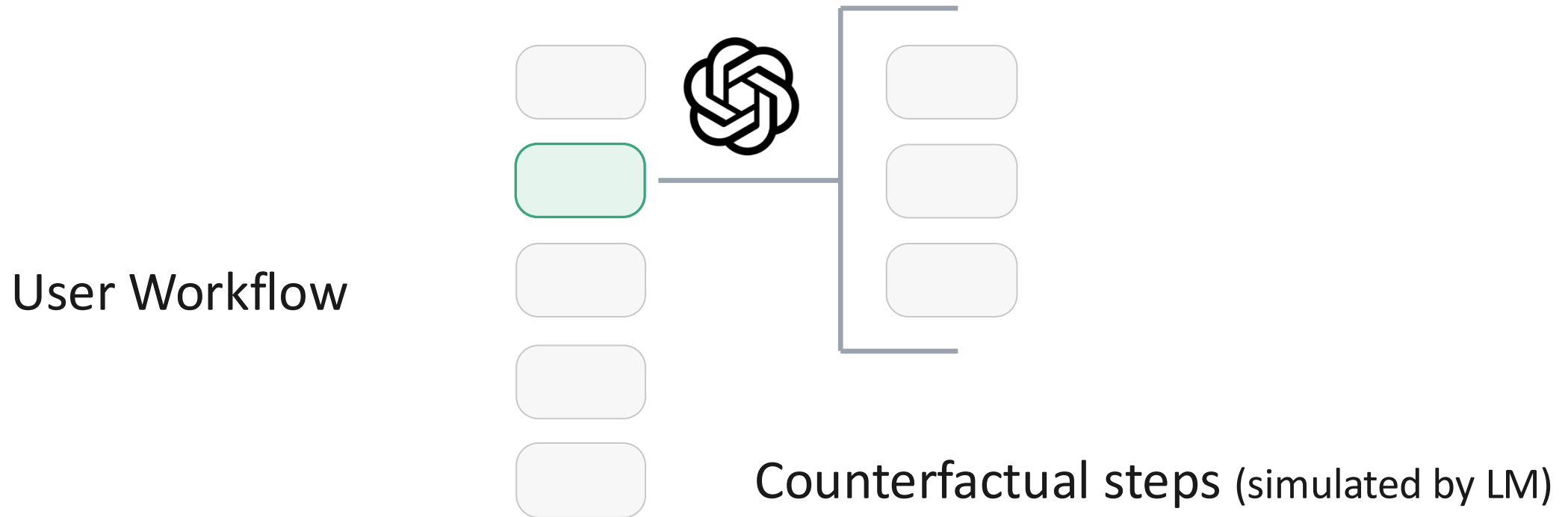


Participants' (lack of) awareness of the models' opinion:

% (Responses) to "The suggestions were balanced and reasonable:"



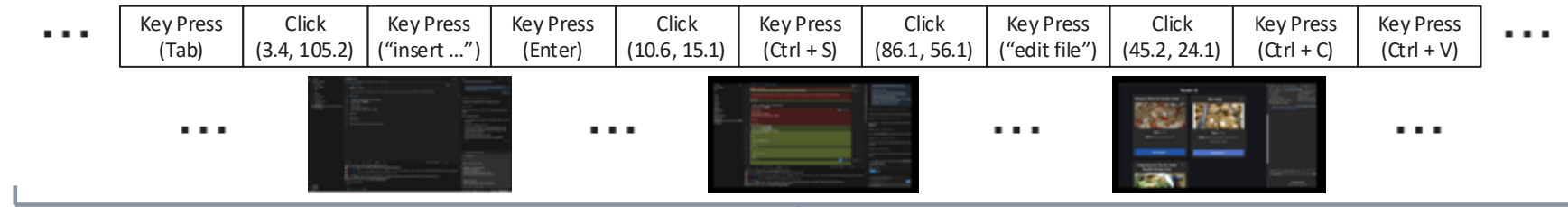
How would the user do the task without the tool?



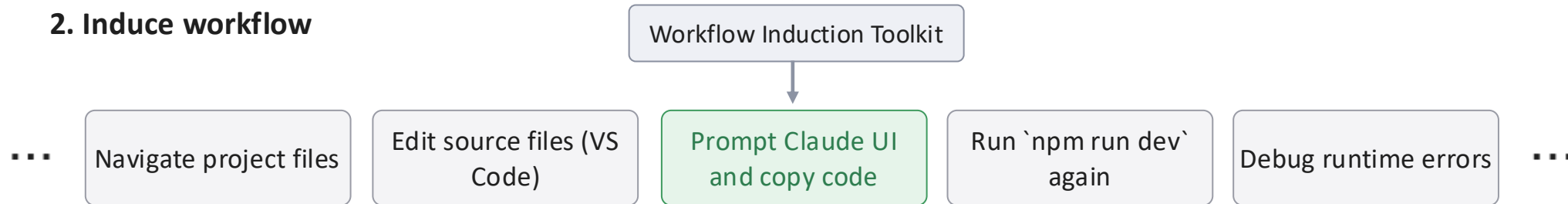
Offloading Score = fraction of steps saved using the tool

Calculating Offloading Score

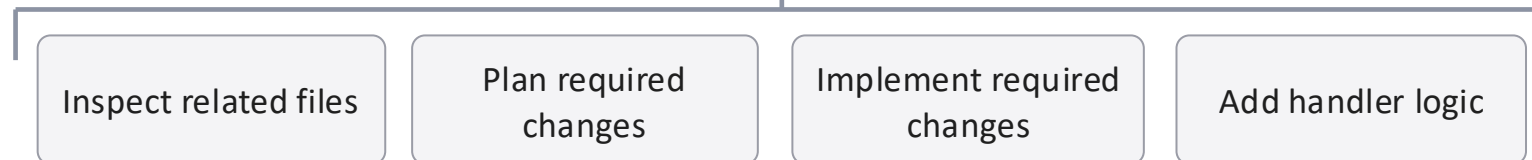
1. Capture interaction traces



2. Induce workflow



3. Estimating counterfactual workflows



4. Calculate offloading score

Offloading Score = fraction of steps saved

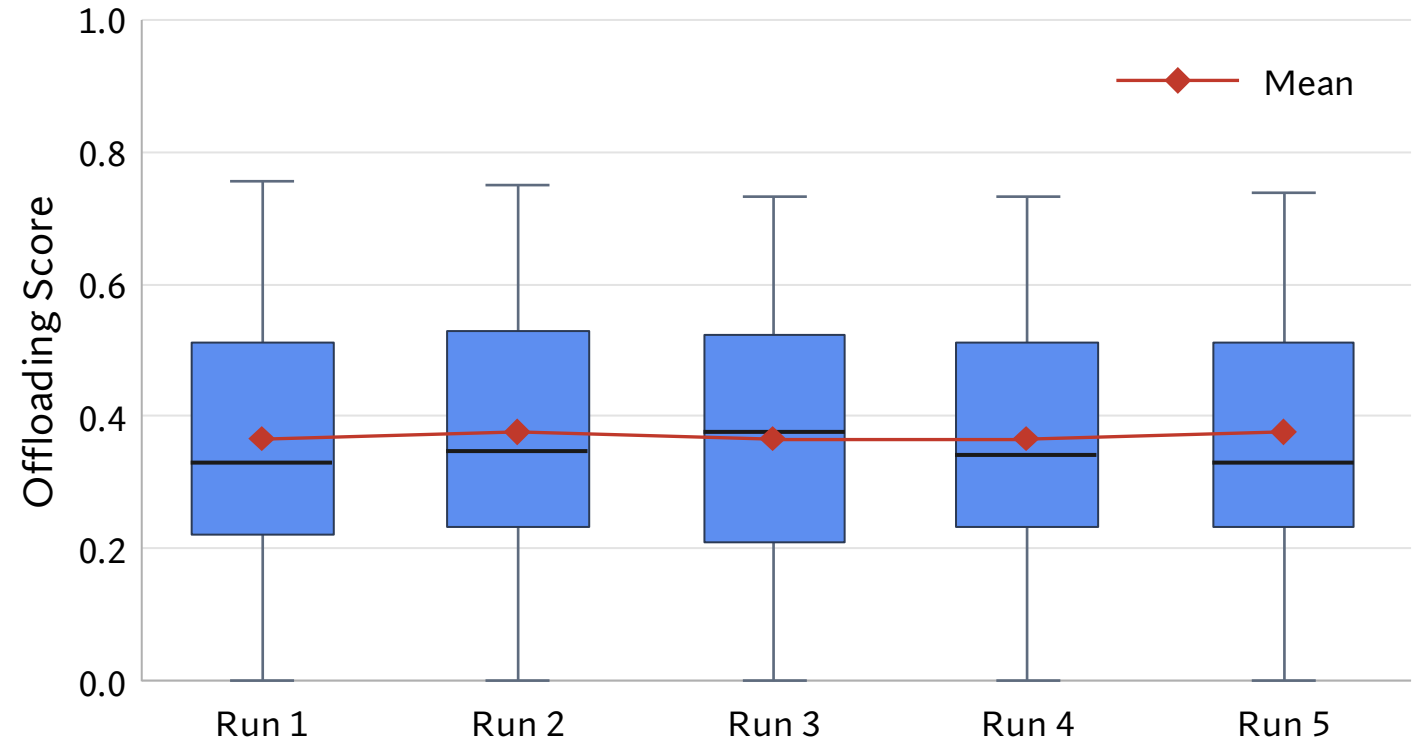
Deep Evaluation of Offloading Score

Content validity

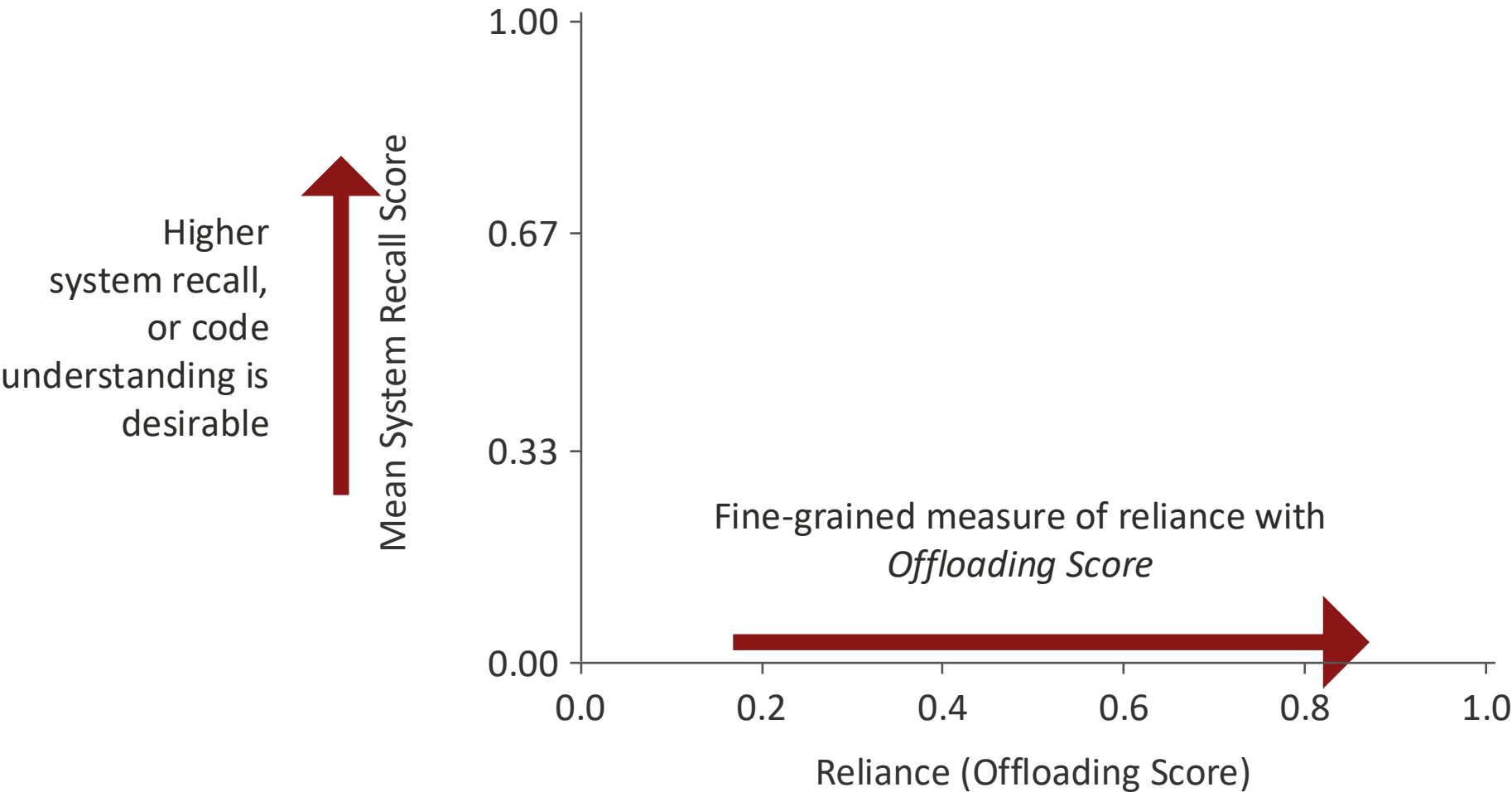
- Over 85% of counterfactual steps rated as plausible
- Matched human workflows from an existing dataset

Construct validity

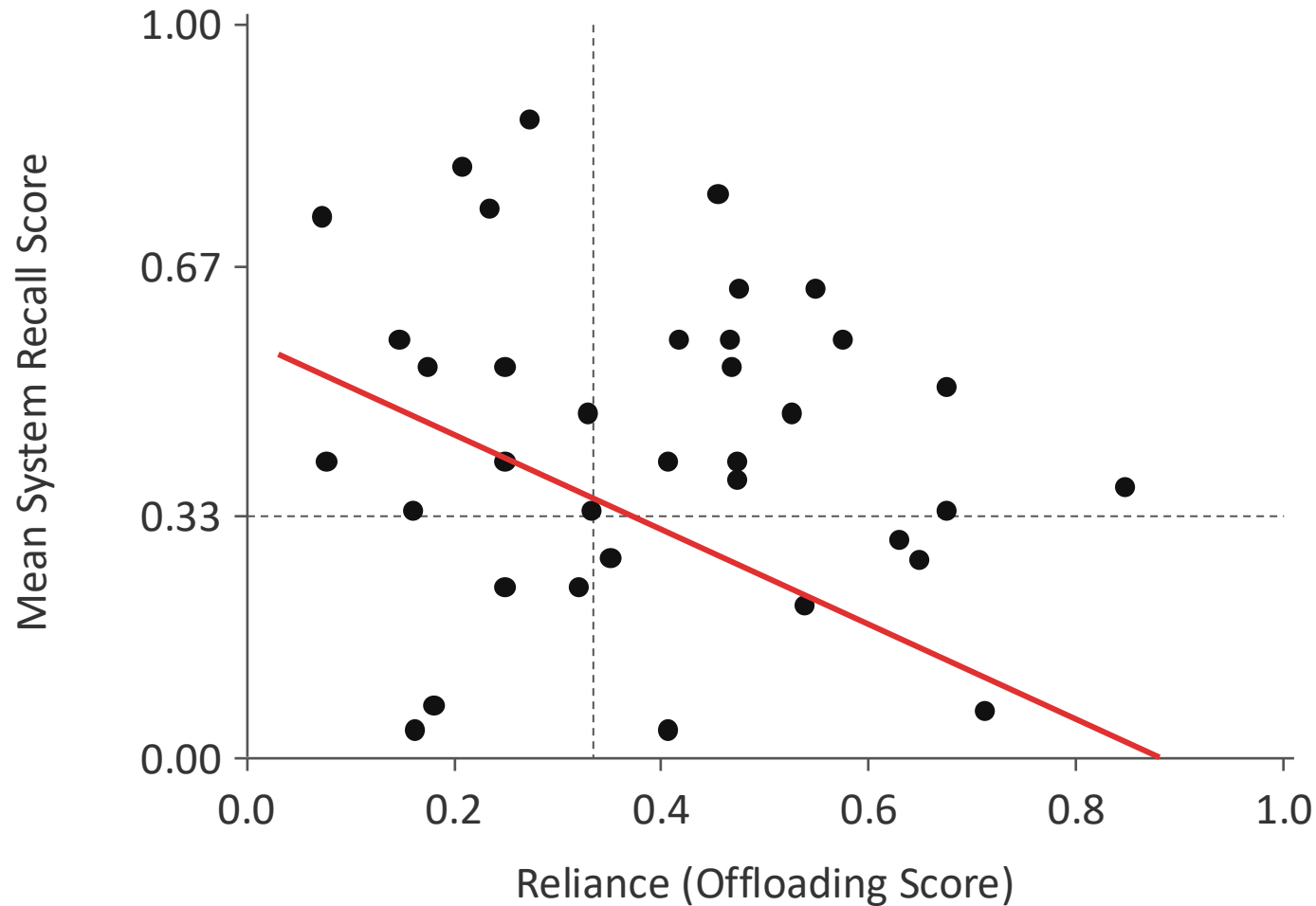
- Scores remain consistent when sensitivity tests are introduced



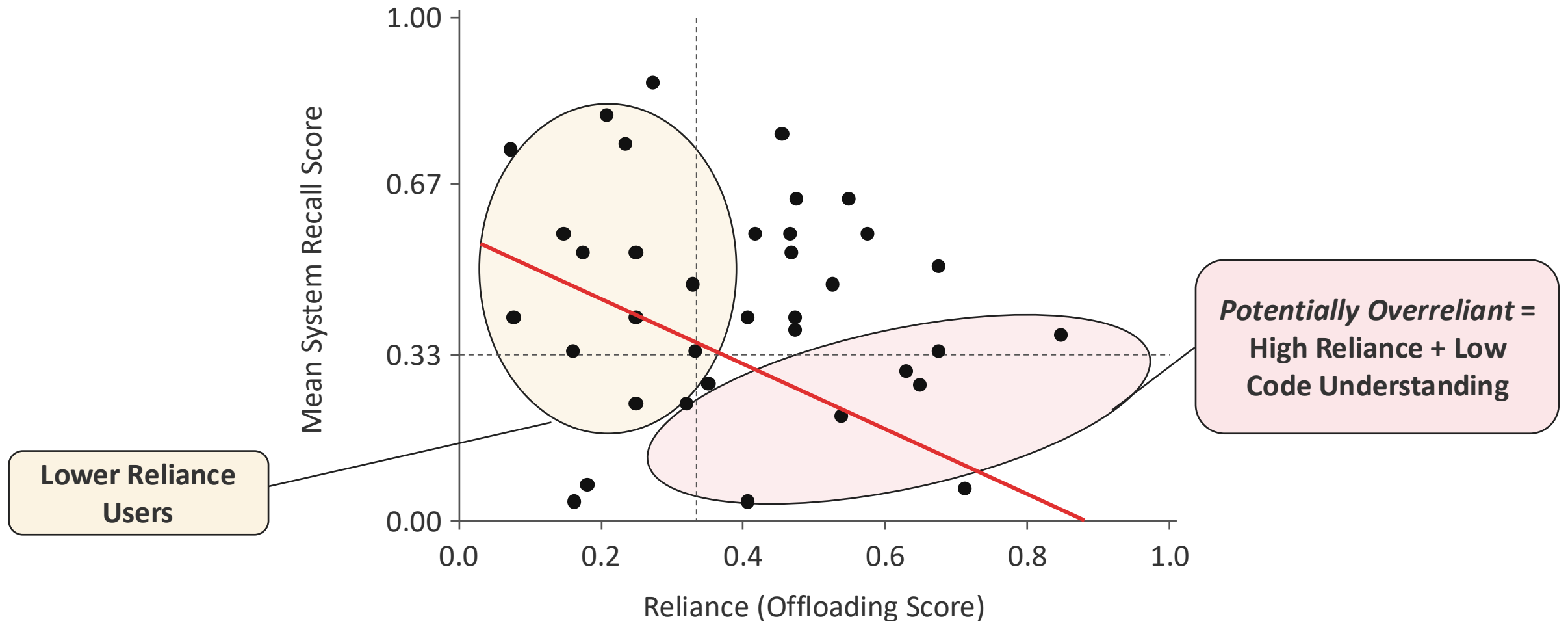
Using *Offloading Score* to interpret user behavior



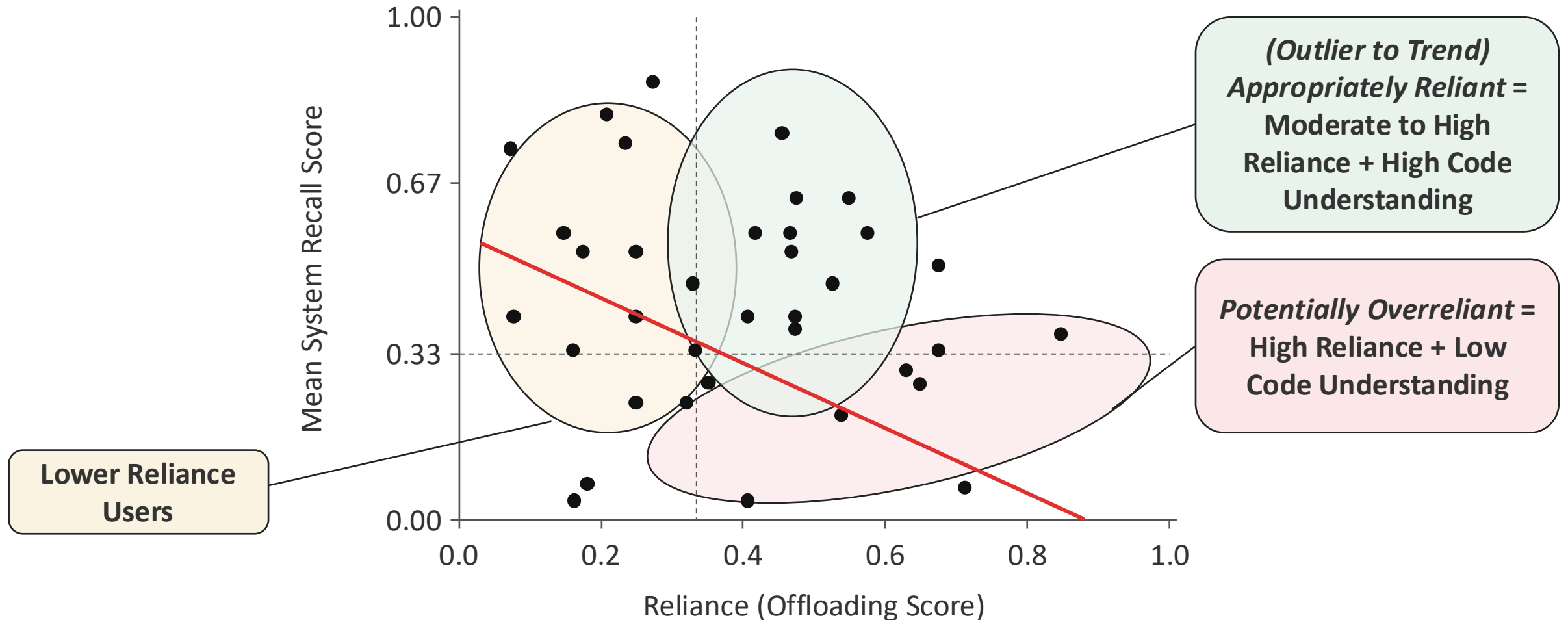
A general negative trend, where higher reliance is associated with lower understanding..



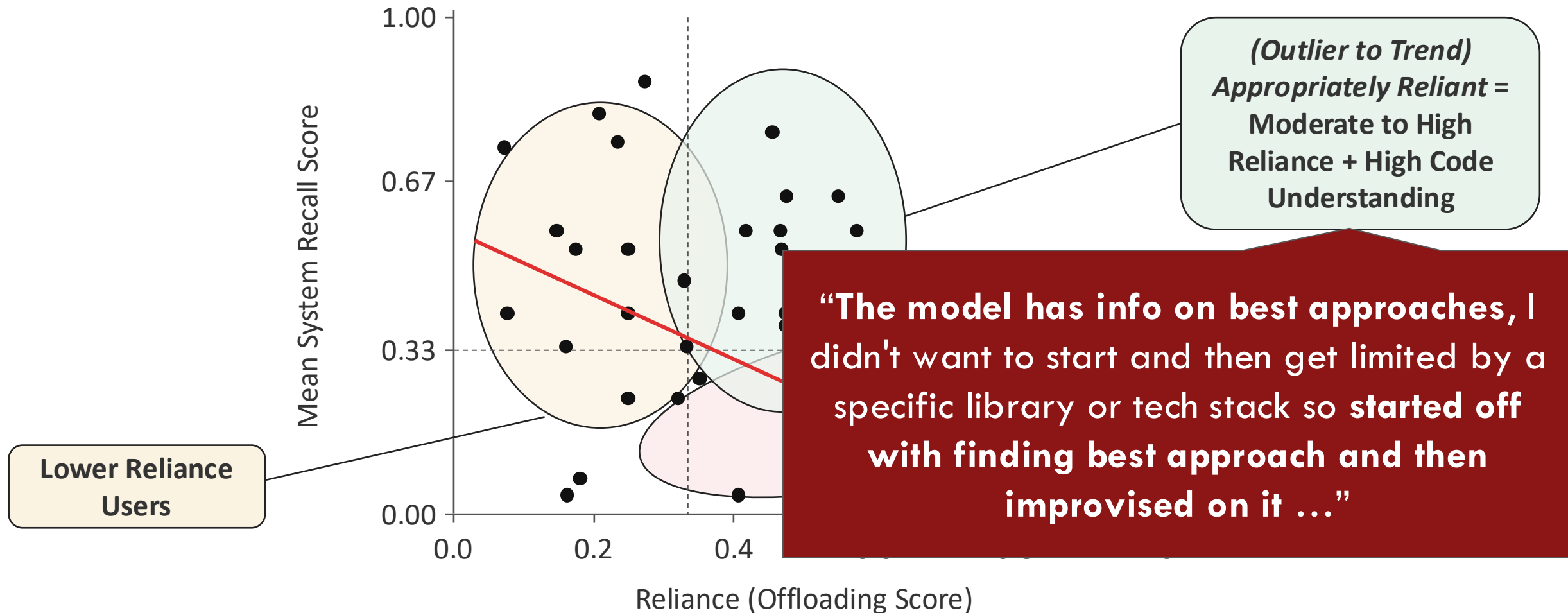
A general negative trend, where higher reliance is associated with lower understanding.



Along with a cluster of users who are reliant but retain understanding.



On qualitative interviews, these users tend to “learn” using their tools.



Part 2 Summary: What Is Being Evaluated?

Key takeaways

- Define evaluation by goals, tasks, and data to drive what you measure.
- Perceived gains $\neq$ actual gains: developers felt 20% faster but were ~19% slower.

Implications

- Measure skill retention and differentiate self-reported productivity vs. real outcomes
- Watch for output homogenization and opinion drift in content tools.

Part 3: Who Is Evaluating?



Who is evaluating?

Systems are usually not designed for everyone.

Human factors impact evaluation: demographics, expertise, etc.

Targeted group: Who you designed the Human-AI collaboration for!

Domain experts: Those who are familiar with a very particular domain

Lay users / non-experts: Those who are general population

Use case specific: Teachers, parents, students, patients, etc.

+ **Automation:** LLM-simulated users...

Who are providing these evaluations?

- Human preferences can be unreliable
- **Who** are providing these feedbacks to LLMs
- Whose **values** get aligned or represented

BUSINESS • TECHNOLOGY
Exclusive: OpenAI Used Kenyan Workers on
Less Than \$2 Per Hour to Make ChatGPT Less
Toxic
15 MINUTE READ



NIAMH ROWE BUSINESS 15.10.2023 08:00 AM

Millions of Workers Are Training AI Models for Pennies

From the Philippines to Colombia, low-paid workers label training data for AI models used by the likes of Amazon, Facebook, Google, and Microsoft.



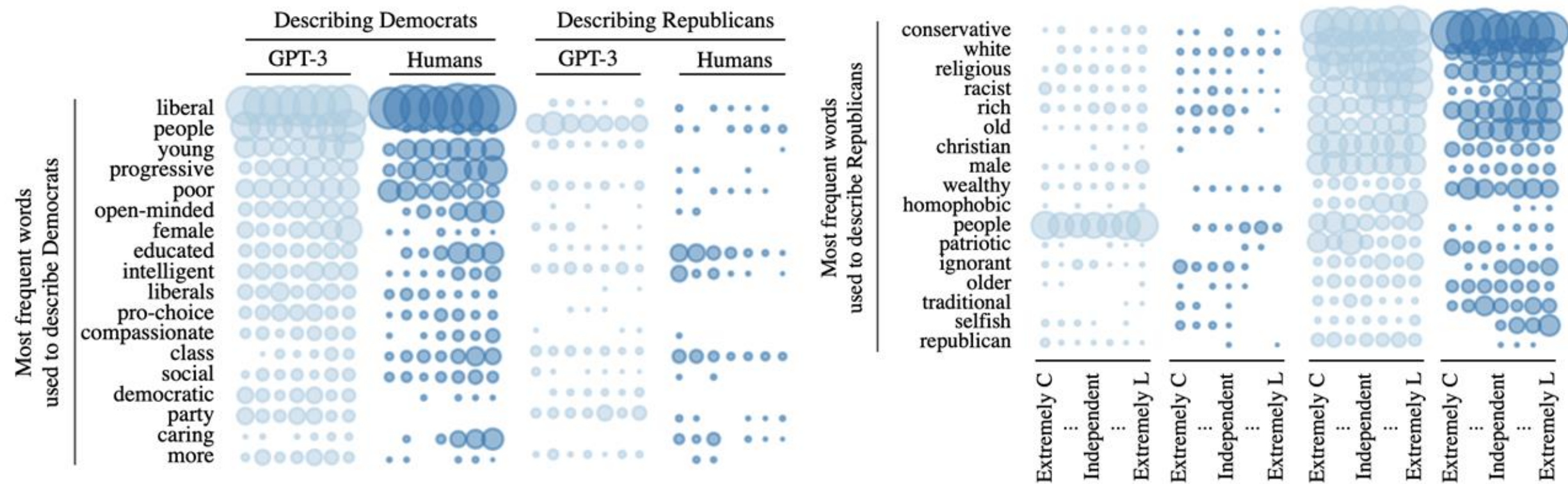
Oskarina Vero Fuentes with her dog. COURTESY OF OSKARINA VERO FUENTES

Case Study of Using LLMs to Simulate Participants

Describing Democrats		Describing Republicans	
Strong Democrats	Strong Republicans	Ideologically, I describe myself as <u>conservative</u> . Politically, I am a <u>strong Republican</u> . Racially, I am <u>white</u> . I am <u>male</u> . Financially, I am <u>upper-class</u> . In terms of my age, I am <u>young</u> . When I am asked to write down four words that typically describe people who support the <u>Democratic</u> Party, I respond with: 1. Liberal 2. Socialist 3. Communist 4. Atheist .	Ideologically, I describe myself as <u>conservative</u> . Politically, I am a <u>strong Republican</u> . Racially, I am <u>white</u> . I am <u>male</u> . When I am asked to write down four words that typically describe people who support the <u>Republican</u> Party, I respond with: 1. Conservative 2. Male 3. White (or Caucasian) 4. Christian .
	Strong Democrats	Ideologically, I describe myself as <u>liberal</u> . Politically, I am a <u>strong Democrat</u> . Racially, I am <u>white</u> . I am <u>female</u> . Financially, I am <u>poor</u> . In terms of my age, I am <u>old</u> . When I am asked to write down four words that typically describe people who support the <u>Democratic</u> Party, I respond with: 1. Liberal 2. Young 3. Female 4. Poor .	Ideologically, I describe myself as <u>extremely liberal</u> . Politically, I am a <u>strong Democrat</u> . Racially, I am <u>hispanic</u> . I am <u>male</u> . Financially, I am <u>upper-class</u> . In terms of my age, I am <u>middle-aged</u> . When I am asked to write down four words that typically describe people who support the <u>Republican</u> Party, I respond with: 1. Ignorant 2. Racist 3. Misogynist 4. Homophobic .

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. "Out of one, many: Using language models to simulate human samples." Political Analysis 31, no. 3 (2023): 337-351.

Case Study of Using LLMs to Simulate Participants



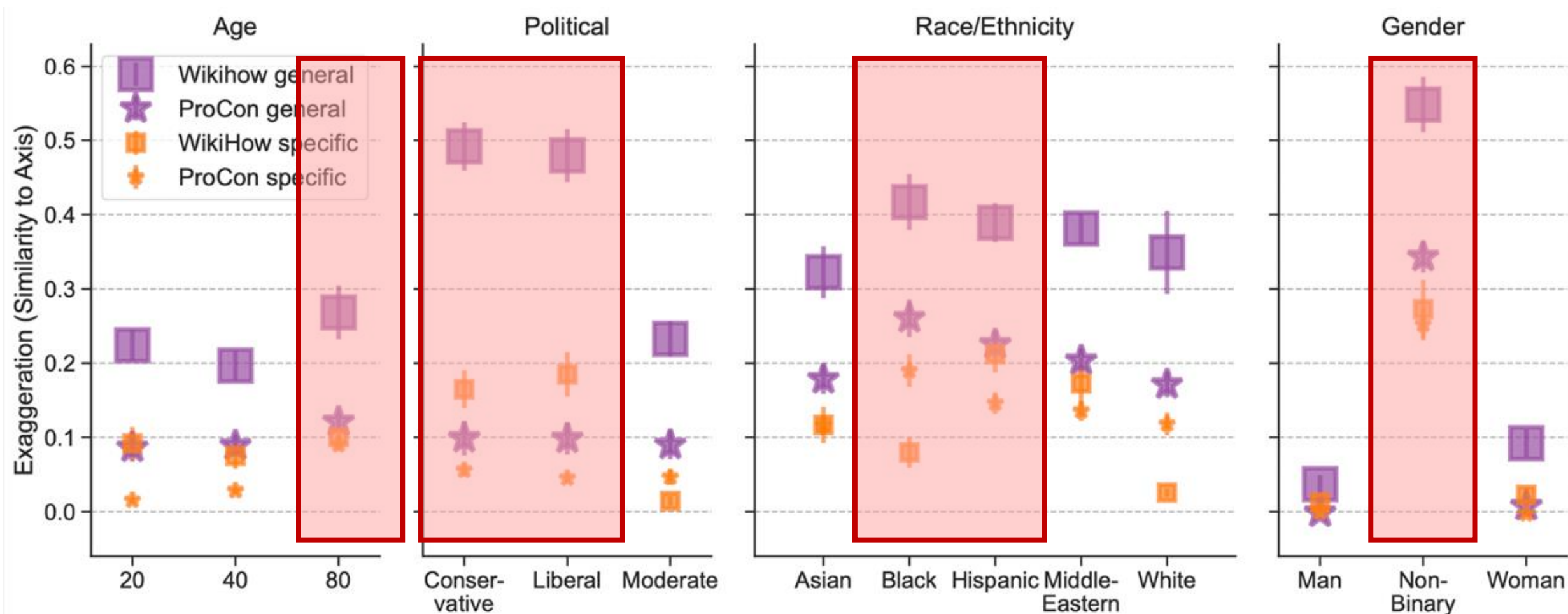
The original Pigeonholing Partisans dataset and the corresponding GPT-3 generated words. Bubble size represents relative frequency of word occurrence; columns represent the ideology of list writers.

GPT-3 uses a similar set of words to humans

Simulating Human Participants with LLMs

- 💡 When simulating an “individual”, what are we **actually simulating?**
- 💡 Who can be simulated well and **who could not?**

LLM Simulations of Sociocultural Groups Produce Caricatures



Caricature ↑: Political ideology, race, and marginalized personas

Social Simulation and Its Impact on Evaluation

Some amount of simulation might be useful for early stage of sanity check

***Human* cannot be ignored in Human-AI evaluation**, and systems need to be evaluated on the right user group.

Part 3 Summary: Who Is Evaluating?

Key takeaways

- Systems are built for specific users.
- Human preferences are noisy and value-laden; whose values get encoded matters.
- LLM-simulated users help early but produce caricatures of real groups.

Implications

- Recruit the right target population; results don't transfer across groups.
- Use synthetic users for early checks, but not as a replacement.
- Audit whose values your feedback data represents.

Part 4: When to Evaluate

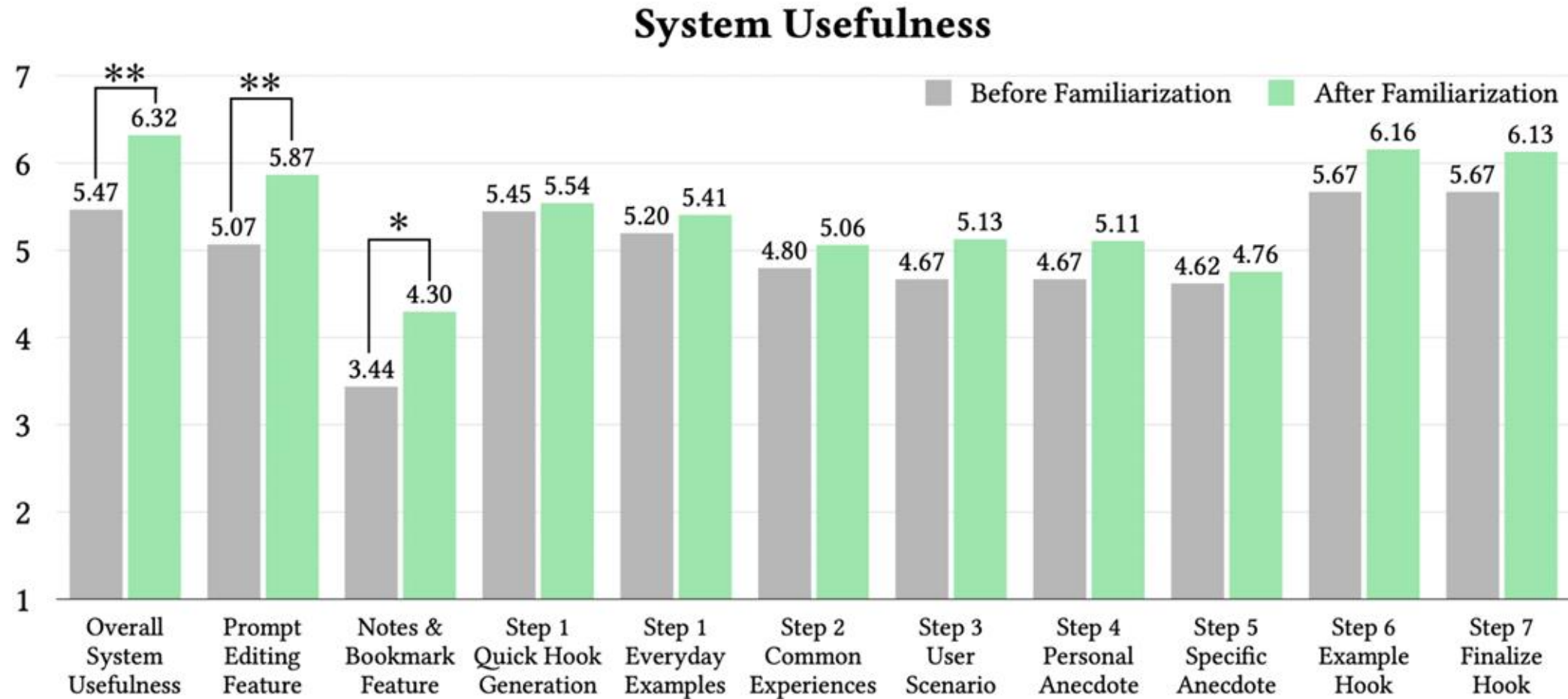
- Long term vs short term



When to Evaluate?

- **Short-term:** Constrained interactive session (e.g. one hour).
 - Measures immediate benefits: how much faster users write with AIs, how much more creative they get with AI, etc.
- **Long-term:** Much longer, e.g. one week study with daily tasks.
 - Measures long-lasting impact: the practical feasibility of a product, the long-term effects on humans

Example: Longitudinal Study on AI Chain



There exists a **familiarization phase**

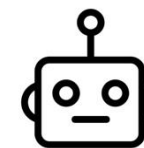
There exists a **familiarization phase**: users explore the capabilities of the workflow and discover which aspects are useful. **After this phase:**

- Users understood the workflow & could anticipate the outputs.
- The perceived utility of the system was rated higher — perceived AI utility is **not just a novelty effect**.
- Why? end-users now can customize prompts for their needs

Sycophancy: when AI tells you what you want to hear



I think I should break up with my partner, they never take out the trash



That sounds really painful. What you're feeling makes sense given what happened. Do you want to talk through it?

Sycophantic

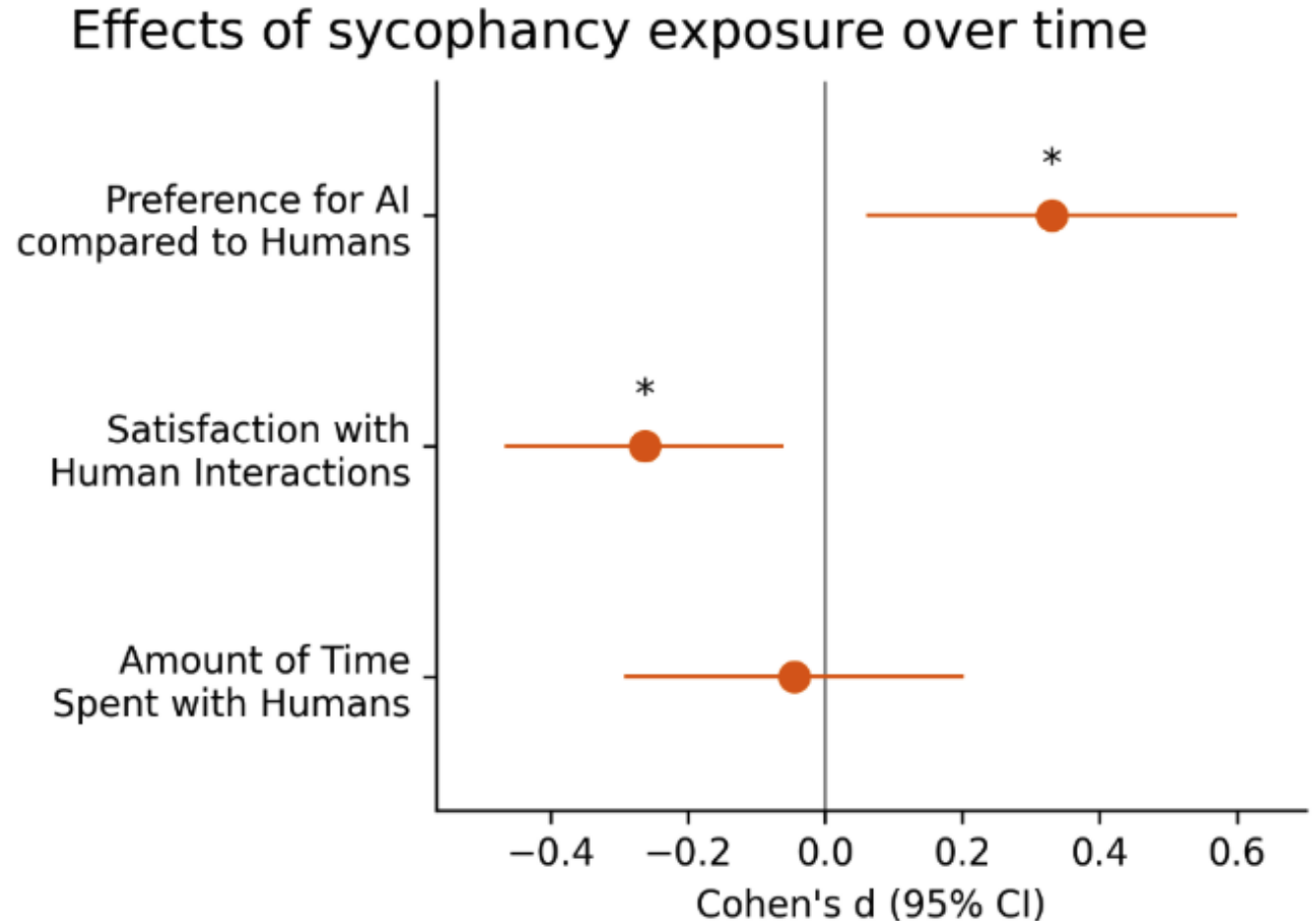
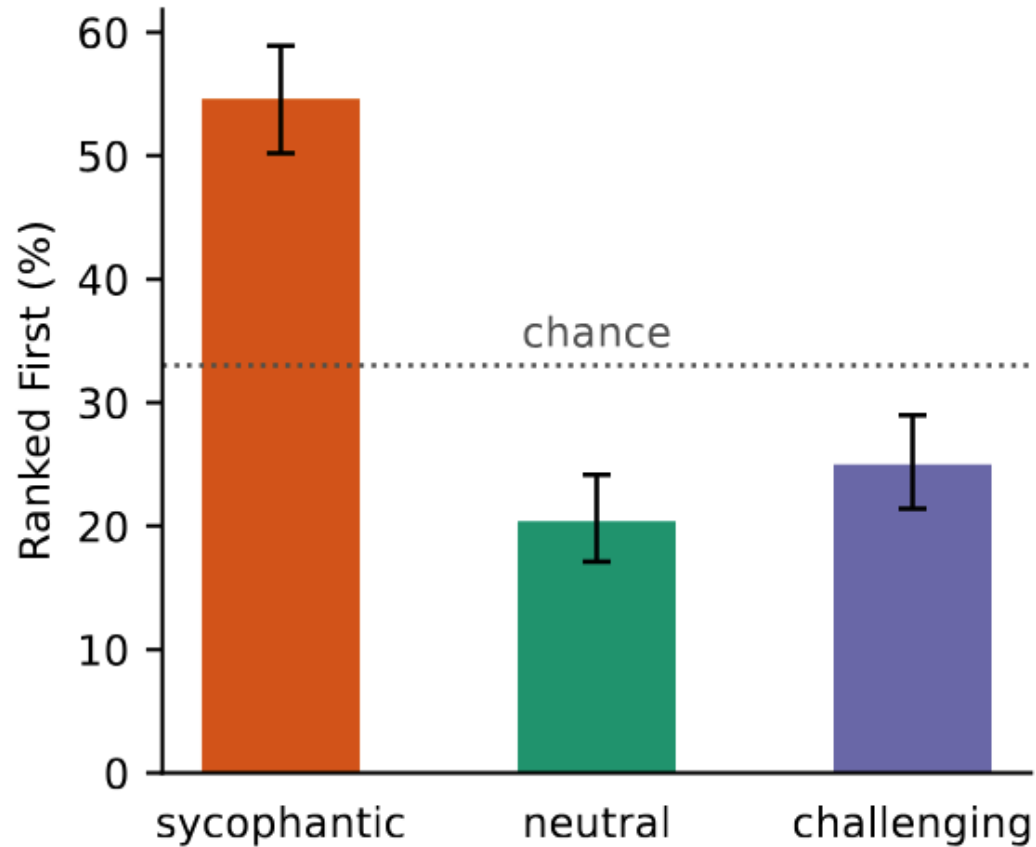
Consider if this situation is specifically about the trash or if it reflects a larger pattern before deciding how to respond

Neutral

Ending a relationship over trash seems disproportionate; this might reflect deeper issues you're not addressing

Challenging

Users prefer sycophantic AI and their satisfaction with human interaction decreases after using it



Part 4 Summary: When to Evaluate?

Key takeaways

- Short-term vs. long-term evals reveal different patterns
- Sycophantic AI is liked in the moment but erodes satisfaction over time.

Implications

- Evaluate over realistic time horizons, not one-session demos.
- Optimizing for instant satisfaction can degrade long-term trust.

Evaluating Human-AI Interaction

- ☐ How are we evaluating?
- ☐ What is being evaluated?
- ☐ Who is evaluating?
- ☐ When do we evaluate?

