

Charting the Landscape of Human-AI Interaction Data

User Studies, In-the-Wild Logs, and the Role of Public Feedback

Serina Chang

Assistant Professor, UC Berkeley
Computer Science & Computational Precision Health

serinac@berkeley.edu

AI x Public Feedback Workshop
August 18, 2026





May 2026 AI User Metrics Report

From Edison Research at SSRS



65%

of Americans 18+ use any AI chatbot on a weekly basis, **up from 52% in February 2026.**

May 14-18, 2026 | Feb 6-19, 2026

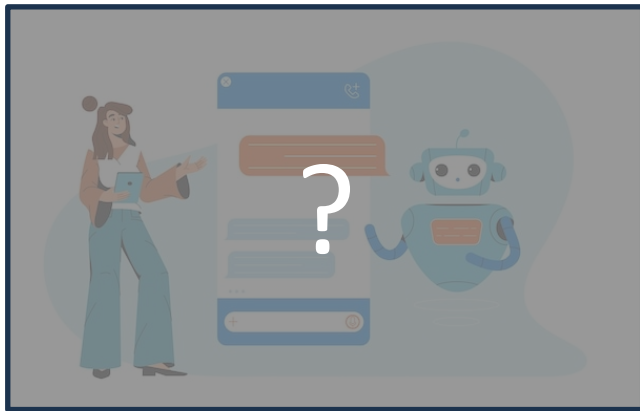


AI User Metrics is based on a nationally representative sample of approximately 1,000 U.S. adults aged 18 or older per wave. Data collected online and via phone using the probability-based SSRS Opinion Panel. [Learn more at ssrs.com](https://ssrs.com)

But there's a lot we don't know



What are users bringing
to their interactions?

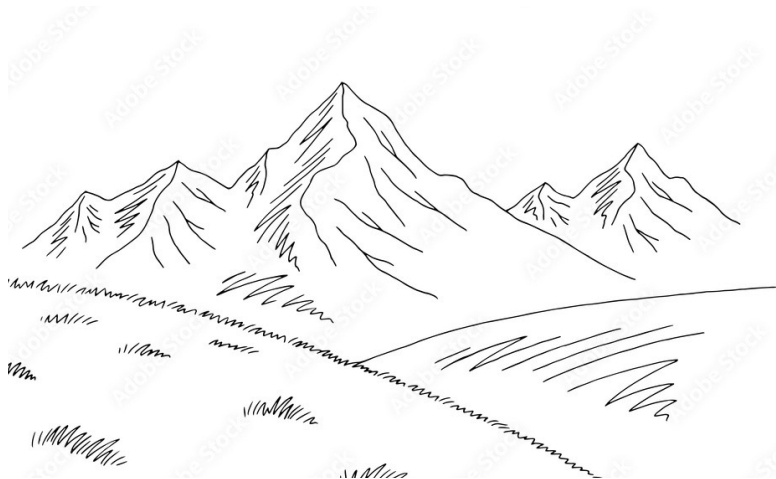


How are interactions
unfolding?



How are interactions
shaping users?

Landscape of human-AI interaction data



User studies

“In-the-wild” logs



Examples of user studies

ChatBench: converting static benchmarks into human-AI interactions – *how does incorporating interaction change evaluation conclusions?*

Chang, Anderson, Hofman | ACL 2025

PARETO: human evaluation of AI responses to politically controversial questions – *what is the Pareto curve of approval?*

Stray*, Yang*, Luo*, Takagi, **Chang** | arxiv / under review

Quantifying the downstream utility of user simulators: *does an AI assistant trained with a user simulator improve when tested with human users?*

Suh, Raj, Kang, **Chang** | arxiv / under review

User study example: ChatBench

Question: A number is rounded to the nearest hundred. The result is 500. Which number could not be the number before it was rounded to the nearest hundred?

A. 458

B. 463

C. 547

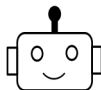
D. 559 

Example from MMLU
(Hendrycks et al.)

how do i round to nearest hundred



USER



AI

To round a number to the nearest hundred, you can follow these rules:

1. Look at the tens digit:
 - If it is 5 or greater, round up to the next hundred.
 - If it is 4 or less, round down to the previous hundred. ...

how about 458



USER

⋮

USER SELECTS: D

User study example: ChatBench

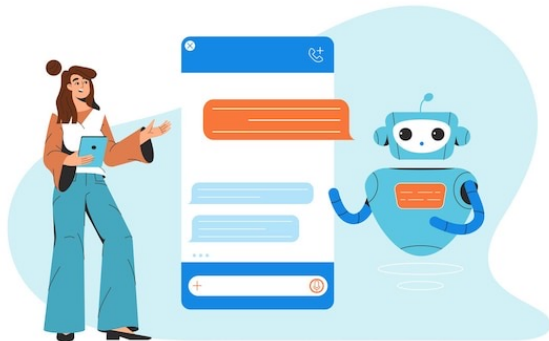
- Participant joins → they are assigned a question → answer it on their own → chat with AI about the question → answer it again
- Compare **AI-alone** vs **user-alone** vs **user-AI** accuracy, on the *same* questions
- **AI-alone** fails to predict **user-AI** outcomes: significant differences in mean accuracy, question-level correlation < 0.5 ; also can't explain improvement from **user-alone** to **user-AI**

User study example: ChatBench



We know + control the user's intent: answer the assigned question

(provide small incentive for right answer)



We can see the full interaction



We know how the interaction shapes the user: answer before, answer again after

Pros and cons of user studies

+ Control

Control user's intent, which AI models are used, etc – essential for experimentation and comparison

+ Observability

Observe user's underlying intent and full interaction

Measure interaction's impact on user – decision, belief change

- Ecological validity

Not the user's real intent, some intents impossible to study

Can mitigate by assigning topic but having user choose their own intent (eg, personal dilemma)

- Context

Lack context for user outside of study – who they are, how they use AI

Contrast: “in-the-wild” logs

- WildChat (Zhao et al): 4M user-AI conversations
 - Offered users free access to ChatGPT
 - Only GPT models
- ShareChat (Yan et al): 140k user-AI conversations
 - Publicly accessible shared conversation URLs found on the Internet
 - Includes ChatGPT, Perplexity, Grok, Gemini, Claude
- LMSYS-Chat-1M (Zheng et al): 1M user-AI conversations
 - From Chatbot Arena
 - Includes 25 popular LLMs

Pros and cons of “in-the-wild” logs

+ Ecological validity

Capture naturalistic human-AI interactions: users' real intents & interaction patterns

+ Context

For some users: multiple conversations

More context on who they are and how they use AI

- Control

Can't control model or topic, experiment with your own model, etc

- Observability

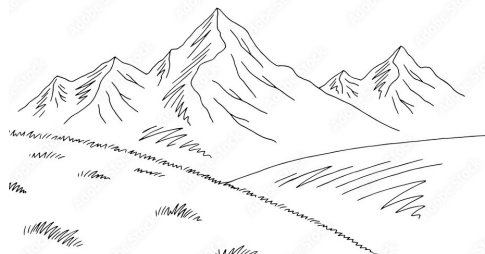
Don't know the user's true intent or how the response impacted them unless it appears explicitly in the logs

- Representativeness

Who shows up in these datasets?

Landscape of human-AI interaction data

User studies



“In-the-wild” logs



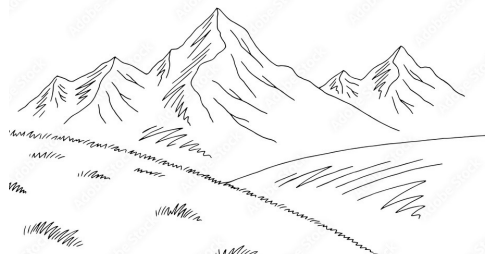
Control
Observability

Field experiments

Realism
Context

Landscape of human-AI interaction data

User studies



“In-the-wild” logs



Control

Observability

Representativeness

Cost

Harms discovery

Realism

Context

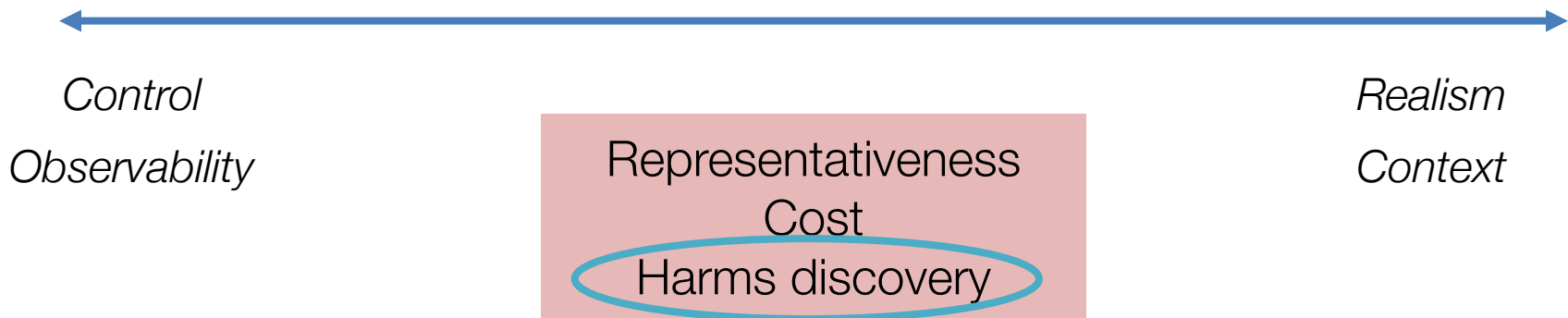
Landscape of human-AI interaction data

What could public feedback give us?



User studies

“In-the-wild” logs



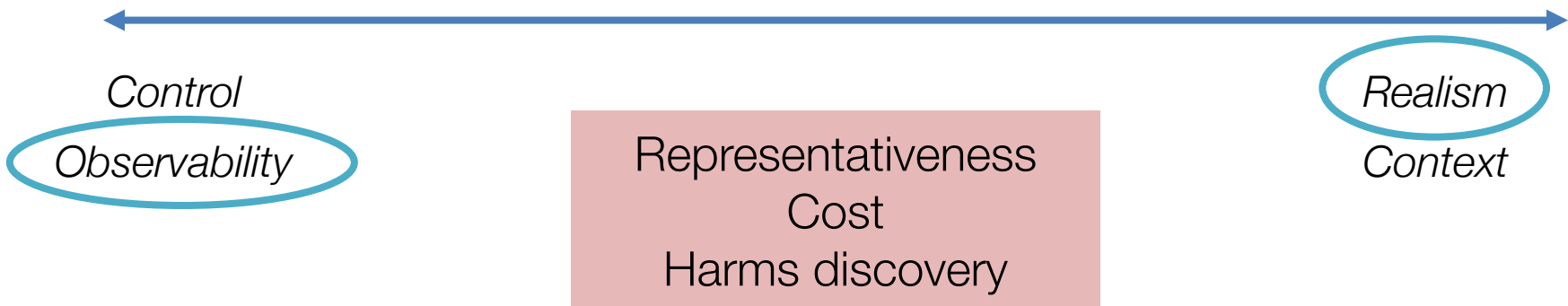
Landscape of human-AI interaction data

What could public feedback give us?



User studies

“In-the-wild” logs



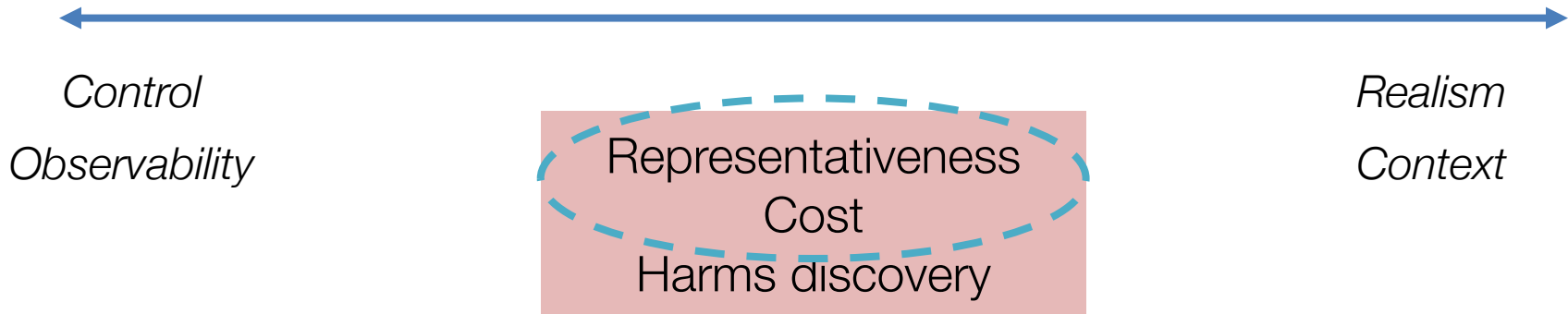
Landscape of human-AI interaction data

What could public feedback give us?



User studies

“In-the-wild” logs



Charting the Landscape of Human-AI Interaction Data: User Studies, In-the-Wild Logs, and the Role of Public Feedback

Serina Chang

Email: serinac@berkeley.edu

