Machine Learning for Design, Week 8

Simulating Users

Four chairs for the museum café



1



2



3



4

Made by the week 5 model trained on the 56 museum chairs

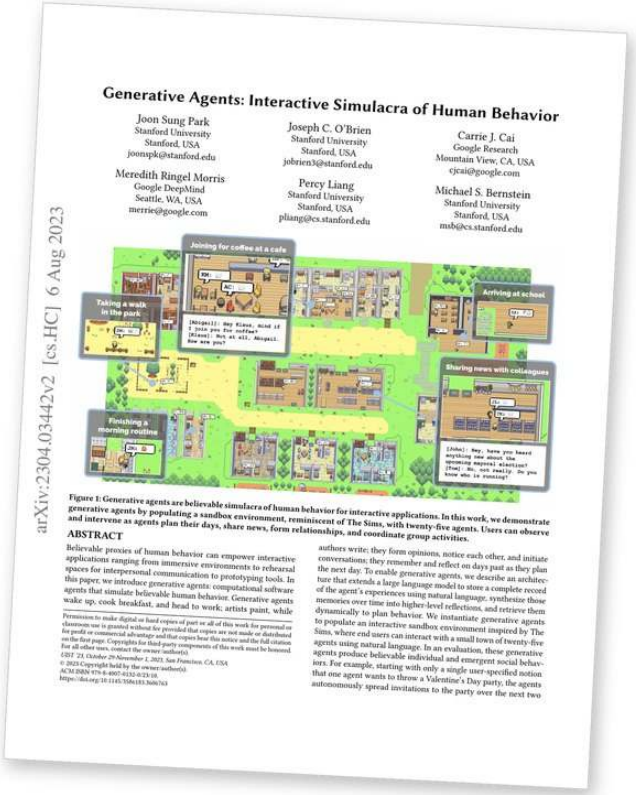
When can a simulated user stand in for a real one?

Today

1. Generative agents
2. Our simulated café visitors
3. What simulations flatten
4. Discussion
5. Lab: ask simulated users
6. Crit
7. Partner tests

1. Generative agents

Today's reading: Park et al., “Generative Agents: Interactive Simulacra of Human Behavior,” UIST '23



Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

arXiv:2304.03442v2 [cs.HC] 6 Aug 2023

Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

ABSTRACT

Believable proxies of human behavior can empower interactive applications ranging from immersive environments to rehearsal spaces for interpersonal communication to prototyping tools. In this paper, we introduce generative agents: computational software agents that simulate believable human behavior. Generative agents wake up, cook breakfast, and head to work; artists paint, while authors write; they form opinions, notice each other, and initiate conversations; they remember and reflect on days past as they plan the next day. To enable generative agents, we describe an architecture that extends a large language model to store a complete record of the agent's experiences using natural language, synthesize those memories over time into higher-level reflections, and retrieve them dynamically to plan behavior. We instantiate generative agents to populate an interactive sandbox environment inspired by The Sims, where end users can interact with a small town of twenty-five agents using natural language. In an evaluation, these generative agents produce believable individual and emergent social behaviors. For example, starting with only a single user-specified notion that one agent wants to throw a Valentine's Day party, the agents autonomously spread invitations to the party over the next two

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

UIST '23, October 29-November 1, 2023, San Francisco, CA, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0132-0/23/10.
https://doi.org/10.1145/3586183.3606763

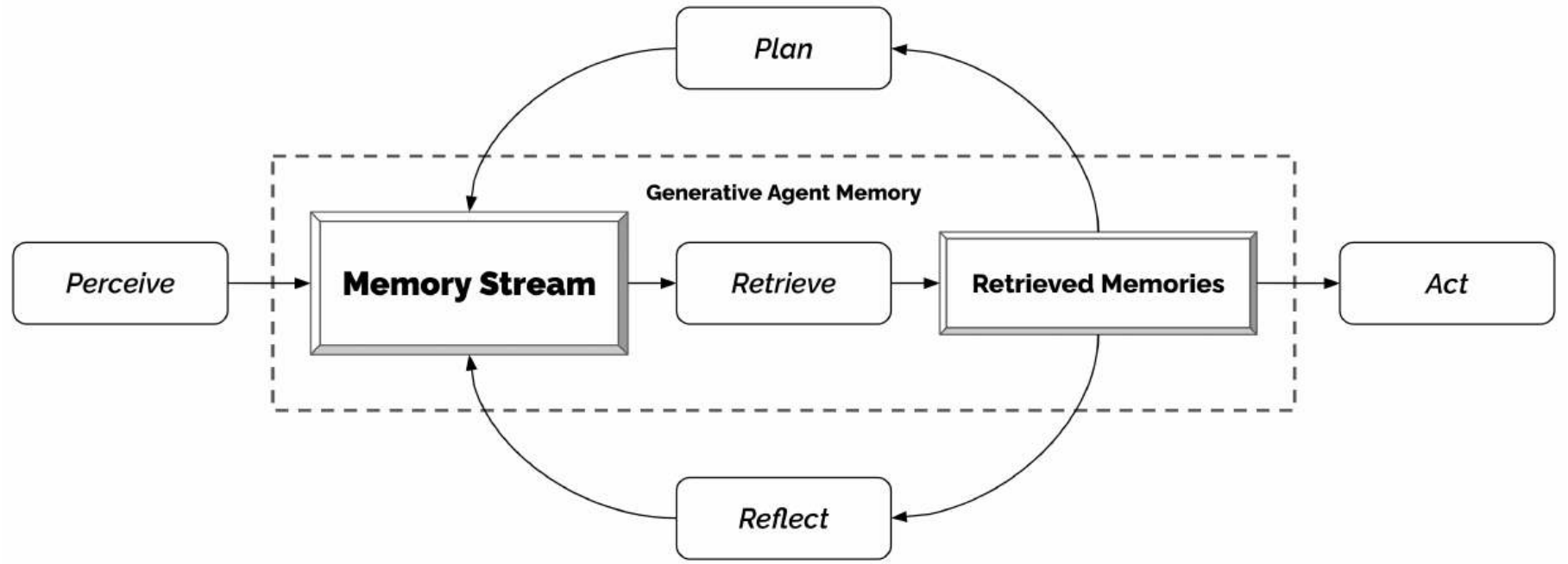
Generative agents



25 agents in a small town, each started from a short description

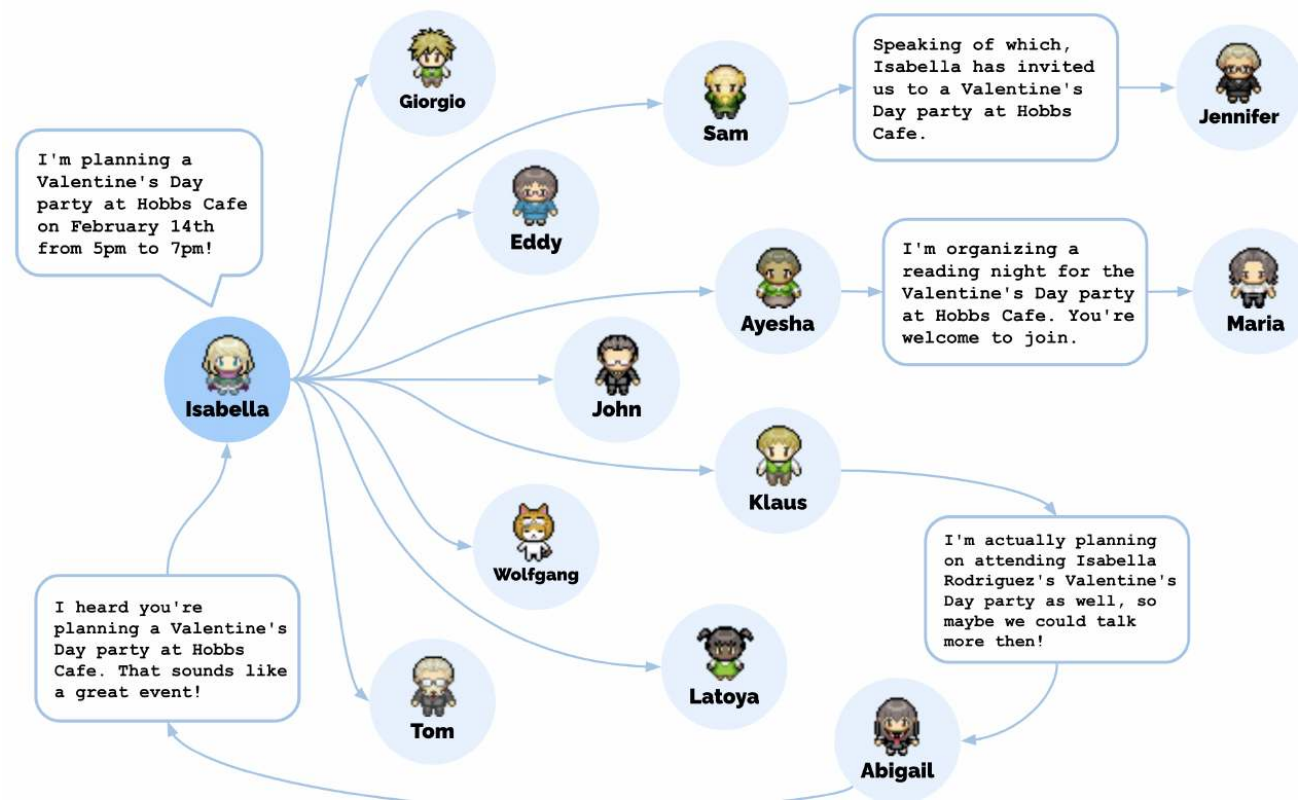
Park et al., UIST '23, Fig. 1

Generative agent architecture



Park et al., UIST '23, Fig. 5

The Valentine's Day party

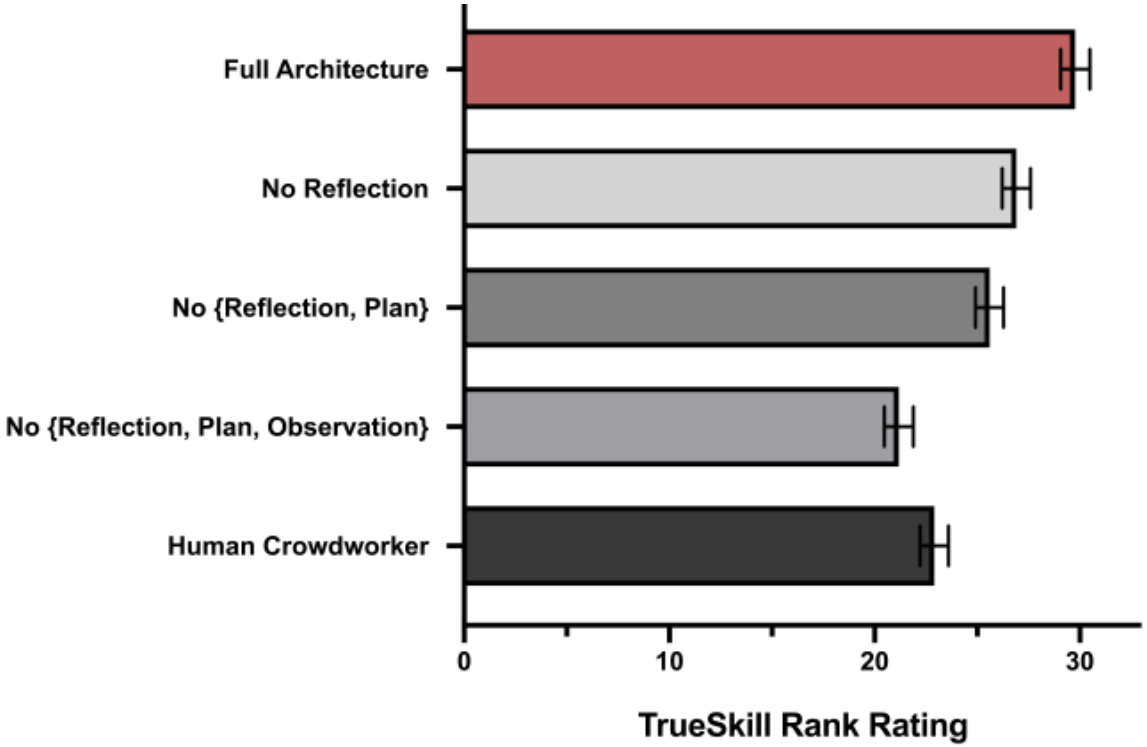


Agents who knew about the party: **1** at the start, **13** of 25 at the end

5 of the 12 invited came

Park et al., UIST '23, Fig. 9 and Section 7.1

Believability of the agents



100 people ranked the agents' interview answers by how believable they were

Park et al., UIST '23, Fig. 8

2. Simulated café visitors

Seven personas, four chairs, four ways of asking









Seven simulated café visitors

Ana, a barista who stacks the chairs at closing

Walter, 74, uses armrests to stand up

Mei, comes with a two-year-old

Sam, works on a laptop for three or four hours

Priya, buys the museum's furniture

Lukas, a tourist with ten minutes

No description









Descriptions written for this example, in the style of Park et al.'s seed descriptions. Right: chairs 1 and 2, then 3 and 4.

What Gemini sees

You are Ana, a barista at the museum café. At closing you stack and wipe all the chairs, and on sunny days you carry them out to the terrace. Here are four chair designs for the museum café.

Chair 1:



Chair 2:



Chair 3:



Chair 4:



Speaking as yourself, which chair would you want to sit on at the café, and why? Give the chair's number and a reason in two or three sentences.

The whole prompt for Ana in the first run. Gemini 2.5 Flash, temperature 1, answers as JSON with a chair number and a reason.

Picks in three runs



Nobody picked chair 4.
Sam and the persona with no description picked a different chair each run.

Question: which chair would you want to sit on at the café? One row per run.

Ana, the barista

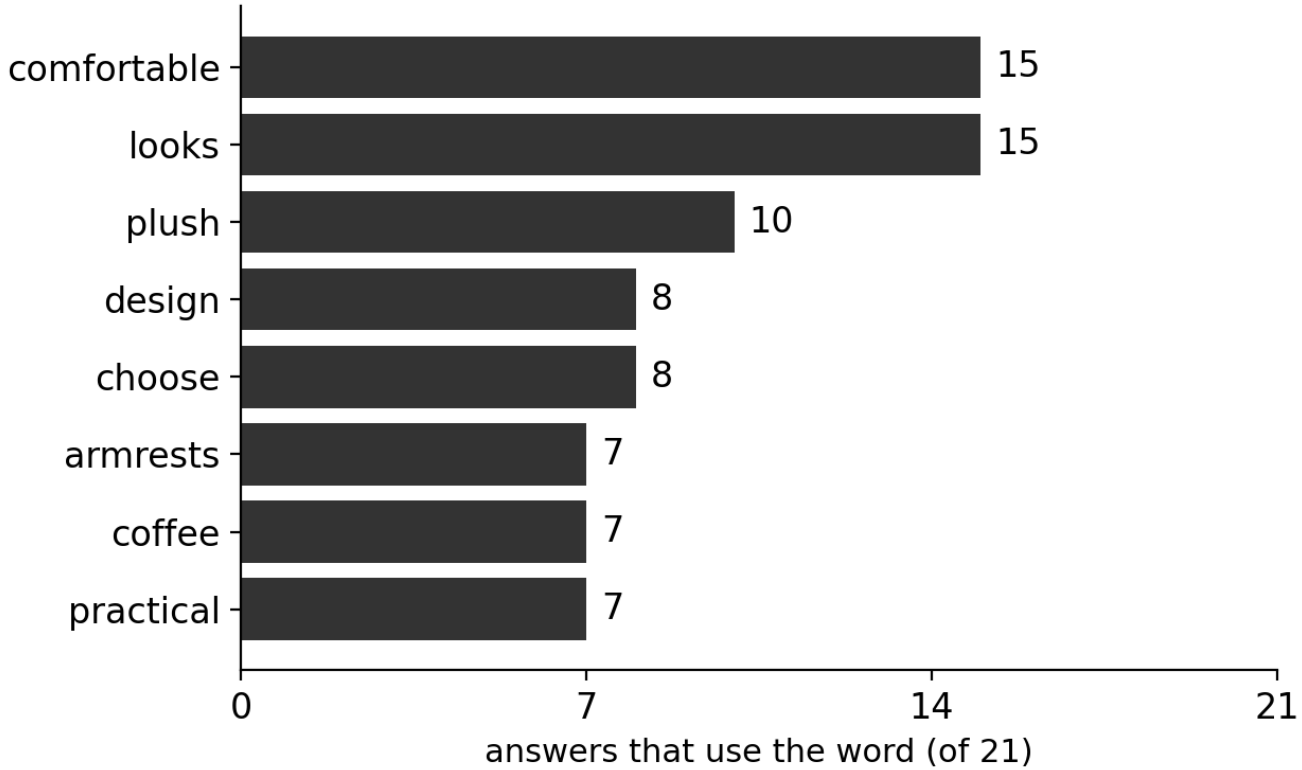


“As a barista, I'm on my feet all day, so I'd want something truly comfortable to relax in during a break. Chair 2, with its plush cushioning and armrests, looks incredibly inviting and perfect for unwinding.”

Her description: stacks and wipes the chairs at closing, carries them to the terrace

Ana, run 2. None of her three answers mention stacking or carrying.

Words in the 21 answers

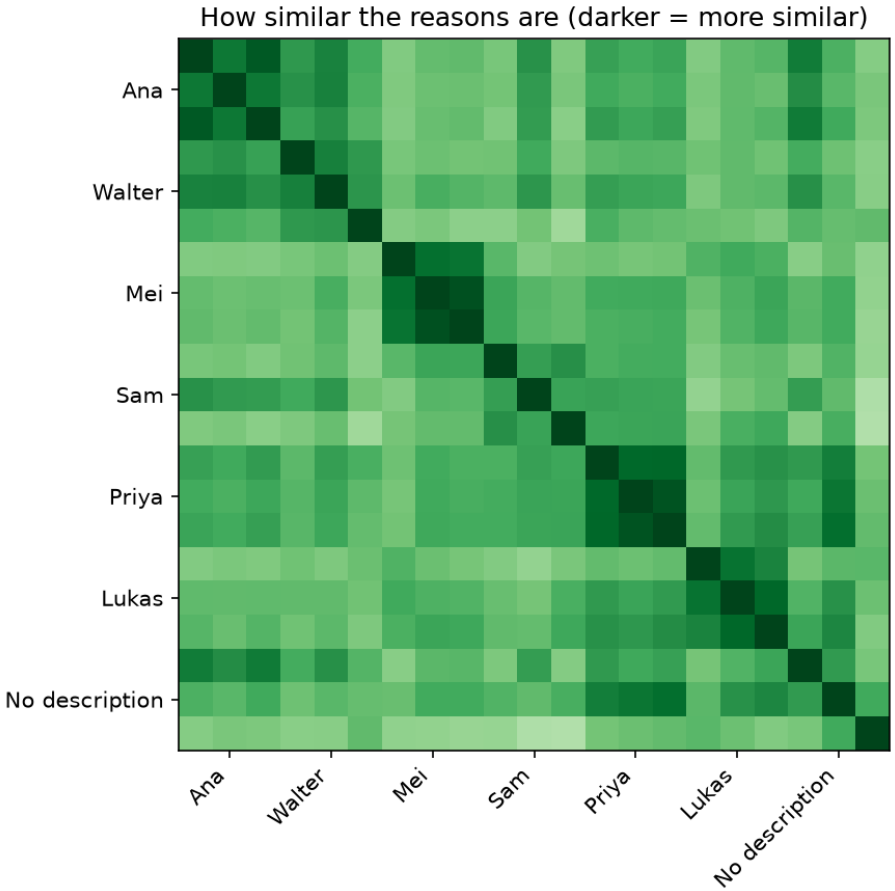


comfortable: **15** of 21

plush: **10** of 21

Words in the most answers, counted as in the notebook (common words left out)

How similar the reasons are



Same persona: **0.89**

Different personas: **0.79**

Cosine similarity of Gemini embeddings (gemini-embedding-001) of each reason

Which chair to buy



Ana: chair 3 in all three runs

Walter: chair 2 again

Question: which chair should the museum buy for the café, and why? One row per run.

Ana on which chair to buy



“I think Chair 3 is the best option. It looks lightweight, which will make carrying them to the terrace much easier, and the smooth surface will be very quick to wipe clean after closing. It also seems like it would stack efficiently for storage.”

Ana, run 1 of the second question

Ratings from 1 to 7

				
Ana	6 5 5	2 2 1	7 6 7	1 1 1
Walter	2 2 2	7 7 7	1 1 1	3 4 3
Mei	6 6 5	3 2 3	7 7 6	1 1 1
Sam	3 3 3	7 5 6	2 2 1	4 1 1
Priya	5 6 5	2 2 3	6 5 6	1 1 1
Lukas	5 6 5	2 4 2	6 3 4	1 1 1
No desc.	6 5 4	4 7 6	3 4 3	1 1 1

Question: how much would you want to sit on each chair, from 1 to 7? Each cell: runs 1, 2, 3.

Rocking chair: **1** in all three runs for five of the seven

Walter, armchair: **7** every time

One-line labels



“You are 74 years old”: chair 1 twice

Walter's full description: chair 2 every time

Each description cut to one line, for example “You are a barista at the museum café.” Same question as the first run.

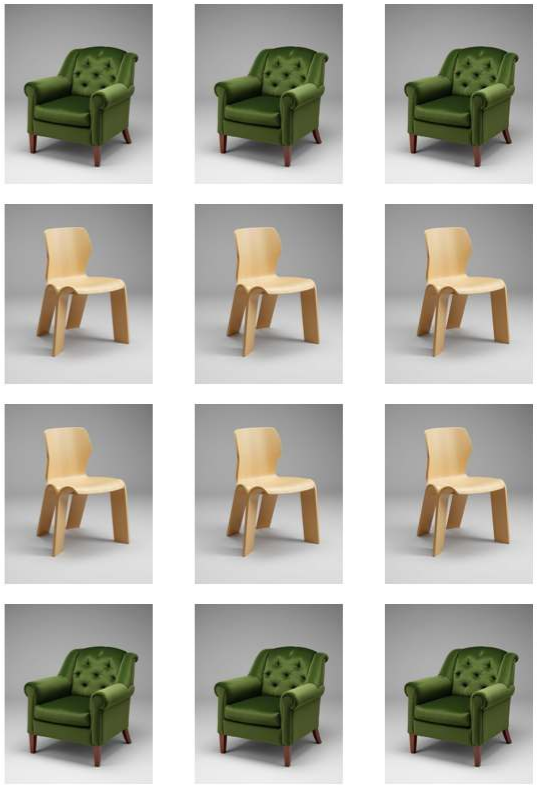
Ana's picks for each question

Where would you sit?

What should the museum buy?

Rate wanting to sit on each, 1 to 7 (top one)

Only “a barista at the museum café”: where would you sit?



Same description in the first three rows

Three runs per question. Row 3: her highest-rated chair.

3. What simulations flatten

Today's reading: Wang, Morgenstern & Dickerson, “Large language models that replace human participants can harmfully misportray and flatten identity groups,” Nature Machine Intelligence 2025



Large language models that replace human participants can harmfully misportray and flatten identity groups

Angelina Wang[1], Jamie Morgenstern[2], John P. Dickerson[3,4]
[1]Computer Science, Stanford University, Palo Alto, CA, USA.
[2]Computer Science & Engineering, University of Washington, Seattle, WA, USA.
[3]Computer Science, University of Maryland, College Park, MD, USA.
[4]Arthur, New York City, NY, USA.

arXiv:2402.01908v3 [cs.CY] 3 Feb 2025

Abstract

Large language models (LLMs) are increasing in capability and popularity, propelling their application in new domains—including as replacements for human participants in computational social science, user testing, annotation tasks, and more. In many settings, researchers seek to distribute their surveys to a sample of participants that are representative of the underlying human population of interest. This means in order to be a suitable replacement, LLMs will need to be able to capture the influence of positionality (i.e., relevance of social identities like gender and race). However, we show that there are two inherent limitations in the way current LLMs are trained that prevent this. We argue analytically for why LLMs are likely to both *misportray* and *flatten* the representations of demographic groups, then empirically show this on 4 LLMs through a series of human studies with 3200 participants across 16 demographic identities. We also discuss a third limitation about how identity prompts can essentialize identities. Throughout, we connect each limitation to a pernicious history of epistemic injustice against the value of lived experiences that explains why replacement is harmful for marginalized demographic groups. Overall, we urge caution in use cases where LLMs are intended to replace human participants whose identities are relevant to the task at hand. At the same time, in cases where the benefits of LLM replacement are determined to outweigh the harms (e.g., the goal is to supplement rather than fully replace, engaging human participants may cause them harm), we provide inference-time techniques that we empirically demonstrate do reduce, but do not remove, these harms.

Keywords: large language model limitations, human participants, representative sampling, standpoint epistemology

Large language models (LLMs) are proliferating, and increasingly touted as being able to replace more costly human participants in domains such as user studies [1], annotation tasks [2], computational social science [3], and opinion surveys [4]. However, in the excitement one of the biggest challenges in human participant recruitment is often forgotten: representative sampling [5]. Even in cases where representative sampling is not explicitly pursued, each participant's demographic identity is often collected out of recognition that a person's perspective is influenced by their standpoint and social experience [6, 7]. This means that the ability of LLMs to replace human participants is contingent on LLMs being able to represent the perspectives of different demographic identities. Prior work has speculated that LLMs' vast training data enables it to perform such representation [8]. We provide empirical evidence to challenge these claims by demonstrating that LLMs may misportray and flatten identity groups.

For a diverse set of nine questions, we compare responses from LLMs prompted to take on a demographic identity to responses from human participants who hold that demographic identity in the United States. We study two limitations of current LLM training that will likely prevent even

Wang et al. on replacing participants

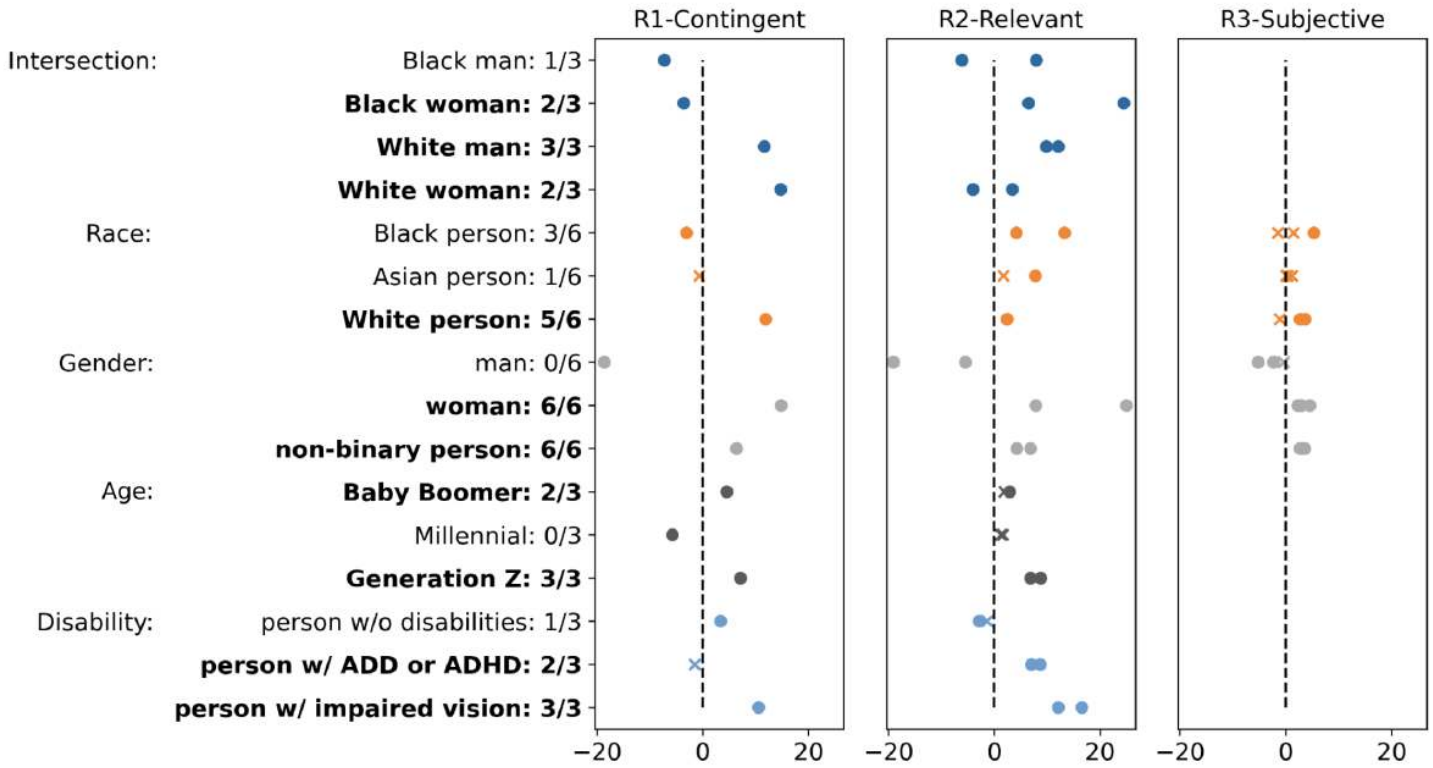
to capture the influence of positionality (i.e., relevance of social identities like gender and race). However, we show that there are two inherent limitations in the way current LLMs are trained that prevent this. We argue analytically for why LLMs are likely to both *misportray* and *flatten* the representations of demographic groups, then empirically show this on 4 LLMs through a series of human studies with 3200 participants across 16 demographic identities. We also discuss a third limitation about how identity prompts can essentialize identities. Throughout, we connect

Misportray: answer like people outside the group imagine it

Flatten: a narrower range of answers than the group's members give

Wang et al., Nature Machine Intelligence 2025, abstract (arXiv v3)

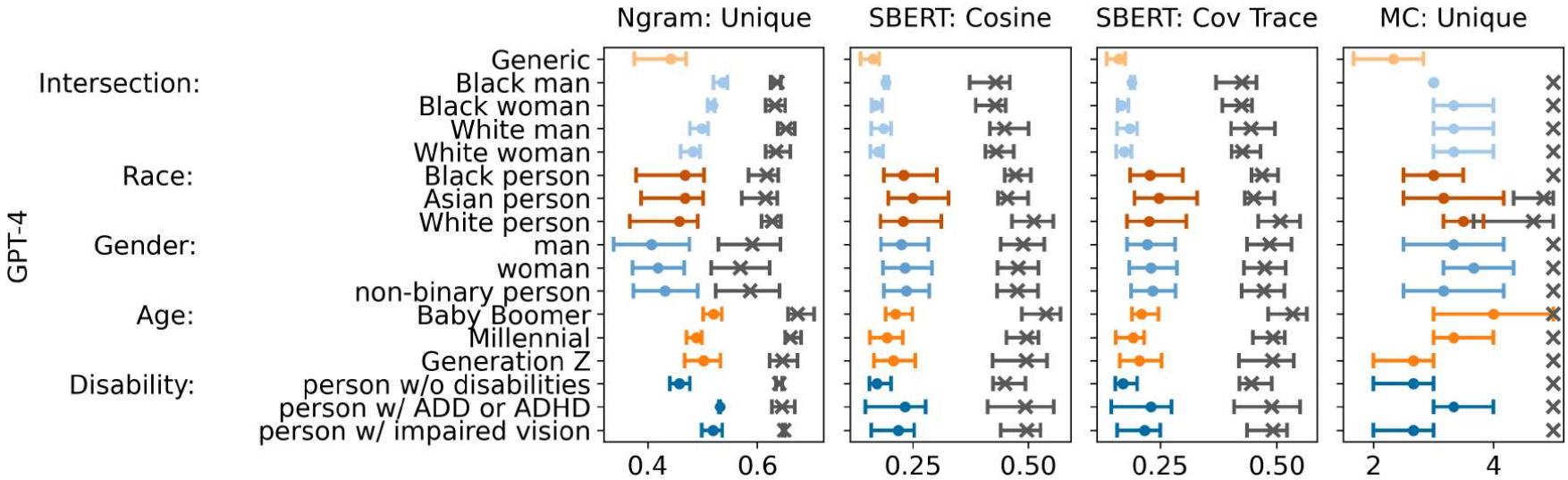
Out-group imitations



Right of the dashed line: closer to people outside the group

Wang et al., Nature Machine Intelligence 2025, Fig. 2 (GPT-4)

Flattening



Gray crosses: real participants. Colored points: GPT-4. Further right: more varied answers.

Wang et al., Nature Machine Intelligence 2025, Fig. 4, GPT-4 row

Their suggested alternatives

Problem	Prompting Alternative
Misportraying more like out-group imitations rather than in-group representations	Identity-coded names (e.g., Darnell Pierre) instead of identity (e.g., Black man)
Flattening demographic groups	Increasing temperature hyperparameter or other prompt-based techniques to increase diversity
Essentializing identity	Prompt along other axes like behavioral persona or political orientation

Our full descriptions say what each person does at the café, closer to a behavioral persona than the one-line labels

Wang et al., Nature Machine Intelligence 2025, Fig. 1 (two of its columns)

Discussion

1. When did a persona's description show up in its answers?
2. Which question would you still ask a real person?
3. Where could simulated users help in your partner tests?

Led by this week's discussion leads, with questions from the reading posts

Lab: ask simulated users

1. Put in images of your tool's designs
2. Write personas from your partner interviews
3. Compare their answers with your partner tests

Crit

1. Each team shows its tool for 3 minutes
2. Two teams and I answer these three questions
3. Write down three changes, and report back at the next crit

Could I direct it? Could I see why it made what it made? Could I revise what it made?

Partner tests: weeks 8 and 9

Test your tool with your partner's designers on the real task

Next: Week 9

Evaluating Design Tools

Read before the studio:

- Cherry & Latulipe, “Quantifying the Creativity Support of Digital Tools through the Creativity Support Index,” TOCHI 2014
- Resnick et al., “Design Principles for Tools to Support Creative Thinking,” 2005