Lecture 3.

Individuals, Groups, and Populations

CS 222: AI Agents and Simulations
Stanford University

Joon Sung Park

Quick housekeeping

Announcements

Enrollment and waitlist information has been sent out! Please let the course staff know if you have not received it.

Note: We have made some edits to the syllabus.

- Assignment 1 will be released *next* week.

Course assistants!

Carolyn

Office hr:
Wednesday 3 - 4 pm
Gates #360

Helena

Office hr:
Monday 9:30-10:30 am
Gates #377

Staff mailing list:

cs222-ai-simulations@cs.stanford.edu

Writing commentaries

Ideal format: 4 ~ 5 paragraphs

- P1: What problems are the two papers trying to tackle? We paired the two papers for each lecture because they offer different perspectives on the general problem space we are studying.
- P2: What approach did the first paper take?
- P23 What approach did the second paper take?
- P4 ~ 5: Discussion — opportunities, limitations, and risks.

Writing commentaries

https://joonspk-research.github.io/cs222-fall24/commentaries.html

Welcome to Week 2 of CS222!

So far: simulations with generative AI offer us a new angle to tackle wicked problems

Lecture 1: Simulations have a rich history, but now there is a new and exciting opportunity.

Lecture 2: Simulations ought to tackle wicked problems.

Course roadmap

- What are the building blocks of simulations?
- How do we create individual agents?
- How do we create the environment?
- How do we establish interactions between agents?
- How do we evaluate the agents?
- How might we envision the language and schema for building simulations?

Assignments

Assignment 1. Creating individual agents

Assignment 2. Creating interactions between agents

In class activity: AgentBank-CS222

Final Project

Quantum unit of simulations

We defined simulations as follows:

A program that defines an *environment* and the behaviors of *individuals*, then outputs the resulting world.

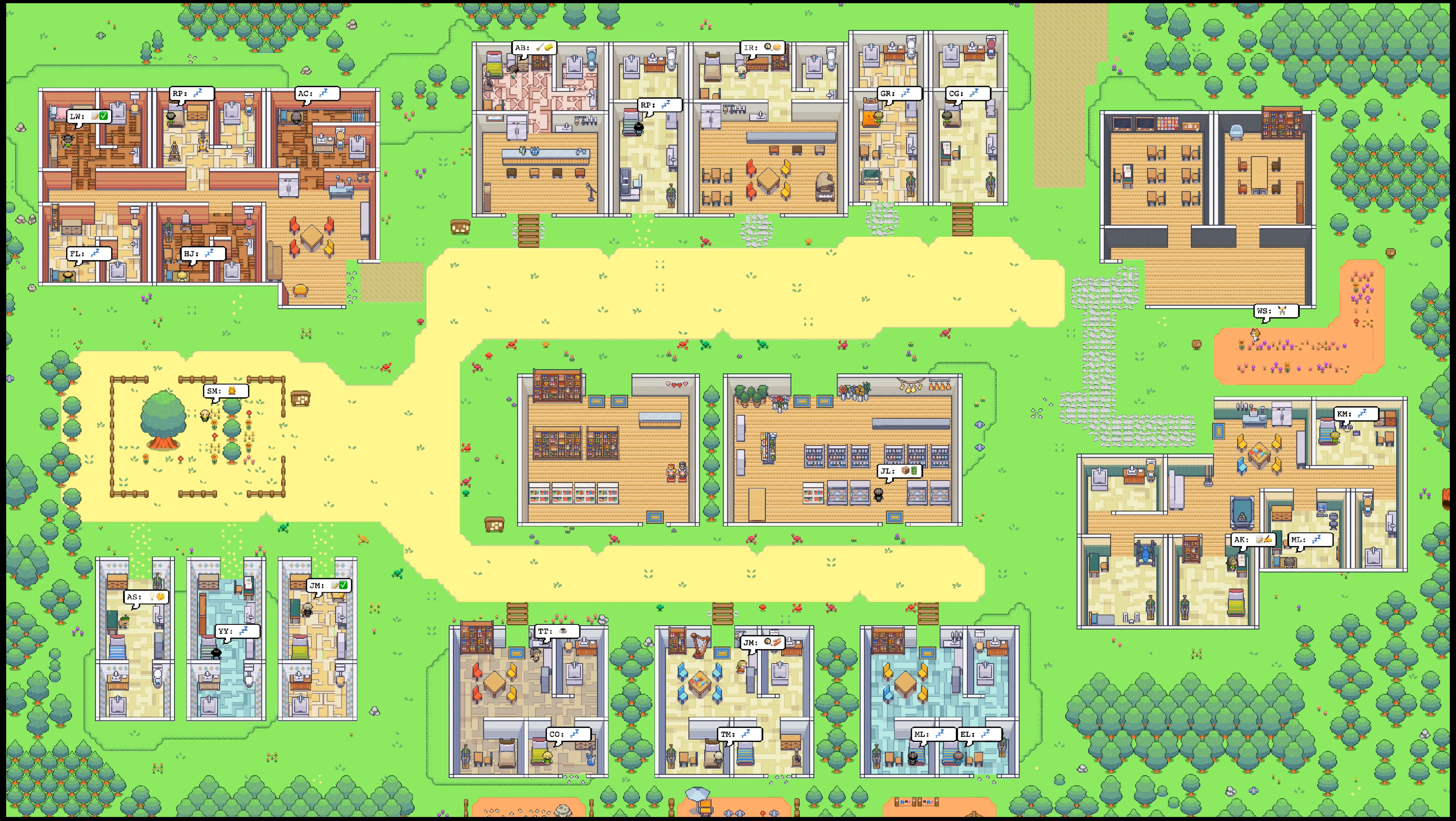
Simulations are a recursive function:

$$W(t) = \left(S_E(t), S_{A1}(t), S_{A2}(t), \ldots, S_{AN}(t)\right)$$

where $W(t+1)$ is recursively defined by the interactions of the environment and agents according to the rules R_E and behaviors $B(A_i)$.

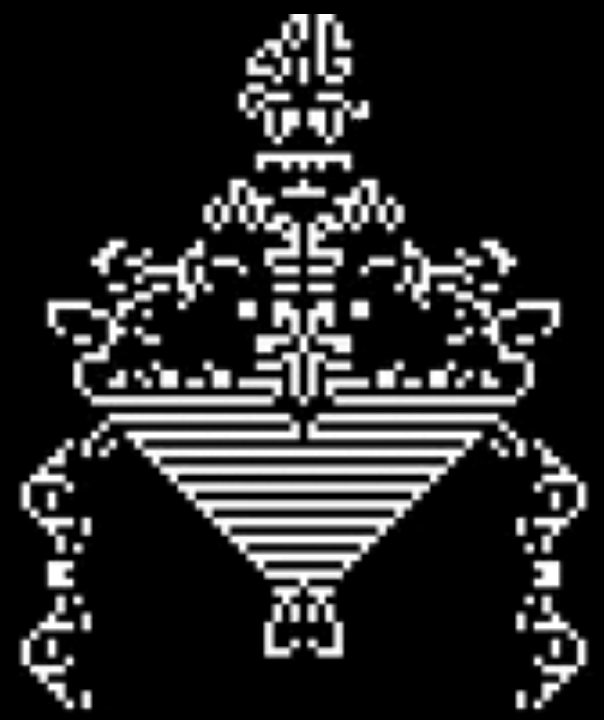
Q: How do we define “individual” agents?

Some simulations offer a perspective on how they define an "individual" agent.



J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023).

But some simulations don’t.



J. von Neumann, Theory of Self-Reproducing Automata, A. W. Burks, Ed. (University of Illinois Press, 1966).

S. Wolfram, A New Kind of Science (Wolfram Media, 2002).

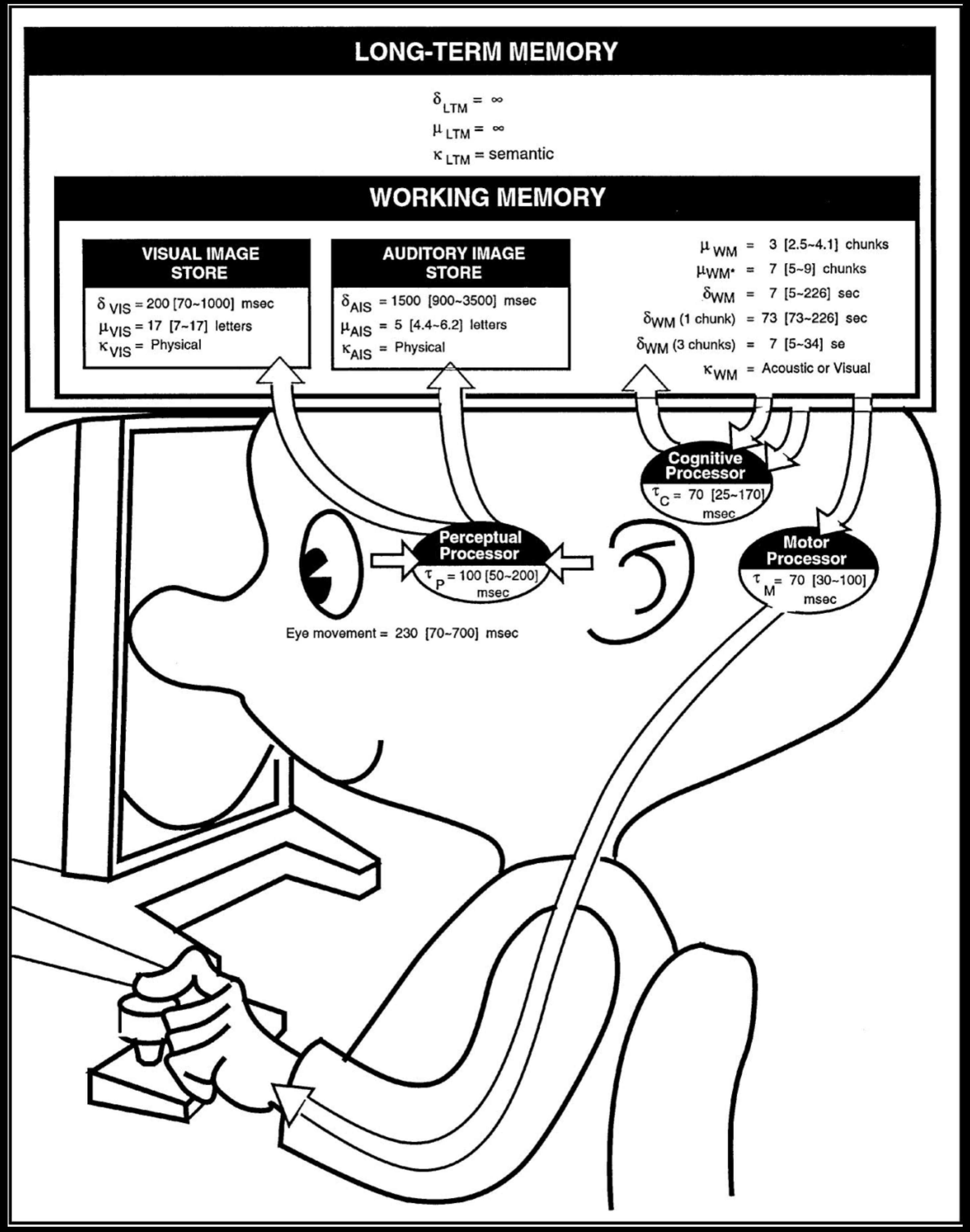
Individuals are the quantum unit of simulations.

Different units of individual agents offer different levels of analysis

Individuals

Groups

Populations


LONG-TERM MEMORY
$\delta_{LTM} = \infty$
$\mu_{LTM} = \infty$
κ_{LTM} = semantic
WORKING MEMORY
VISUAL IMAGE STORE
δ_{VIS} = 200 [70~1000] msec
μ_{VIS} = 17 [7~17] letters
κ_{VIS} = Physical
AUDITORY IMAGE STORE
δ_{AIS} = 1500 [900~3500] msec
μ_{AIS} = 5 [4.4~6.2] letters
κ_{AIS} = Physical
μ_{WM} = 3 [2.5~4.1] chunks
μ_{WM^*} = 7 [5~9] chunks
δ_{WM} = 7 [5~226] sec
δ_{WM} (1 chunk) = 73 [73~226] sec
δ_{WM} (3 chunks) = 7 [5~34] se
κ_{WM} = Acoustic or Visual
Cognitive Processor
τ_C = 70 [25~170] msec
Perceptual Processor
τ_P = 100 [50~200] msec
Motor Processor
τ_M = 70 [30~100] msec
Eye movement = 230 [70~700] msec

The granularity of simulations is often chosen based on practicality and the specific goal at hand.

SK Card, TP Moran, and A Newell. 1983. The psychology of human-computer interaction. (1983).

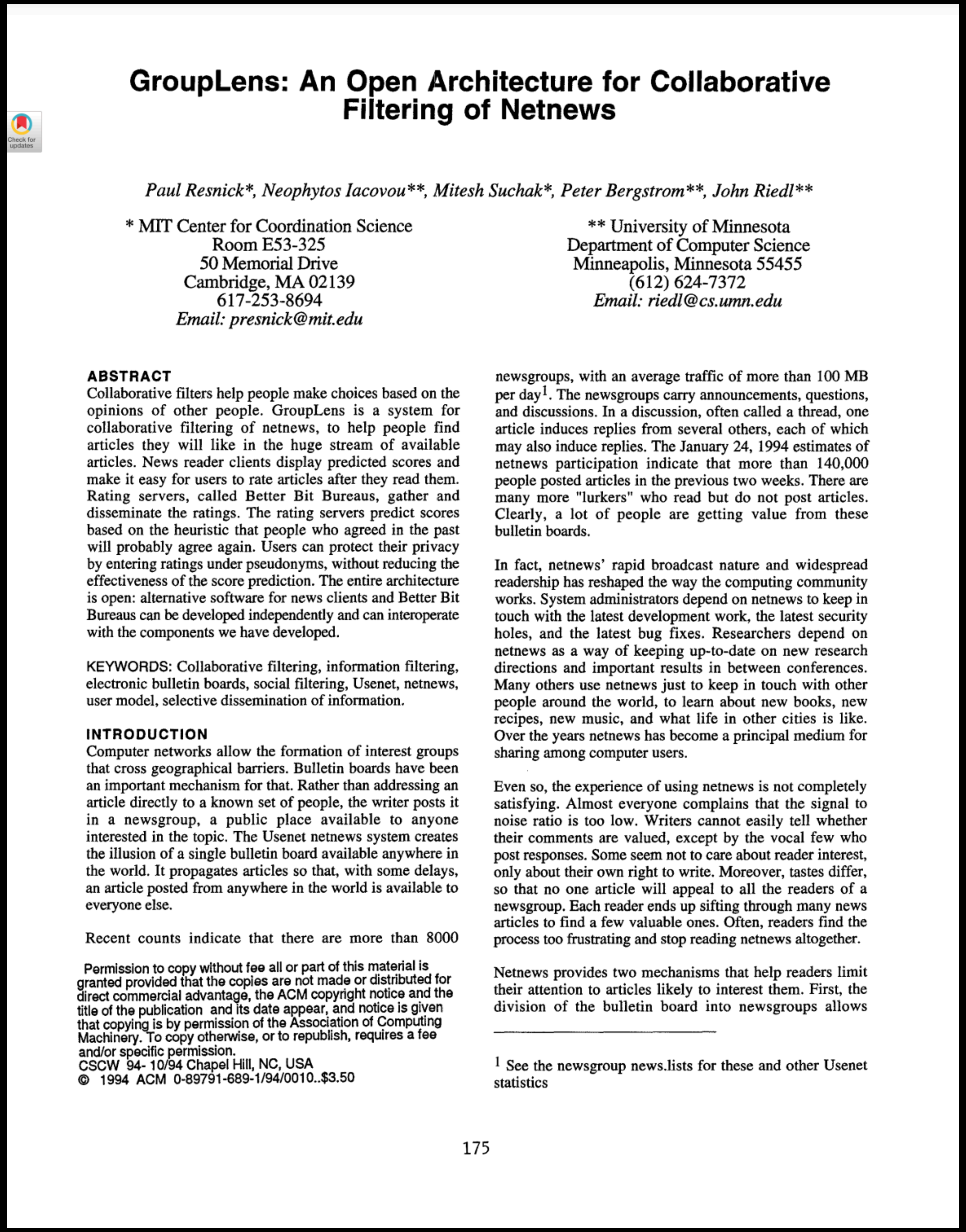
What are individuals?

- An individual is a single person who possesses unique qualities, traits, beliefs, and experiences that distinguish them from others.
  - Personality traits
  - Beliefs and values
  - Appearance
  - Behaviors and expressions

Simulations of individuals allow us to ask highly granular questions that uniquely apply to that one person.

For instance,

- Would this individual like these search results or recommendations?
- How would this individual react to experimental treatments?



GroupLens: An Open Architecture for Collaborative Filtering of Netnews

Paul Resnick, Neophytos Iacovou**, Mitesh Suchak*, Peter Bergstrom**, John Riedl***

* MIT Center for Coordination Science
Room E53-325
50 Memorial Drive
Cambridge, MA 02139
617-253-8694
Email: presnick@mit.edu

** University of Minnesota
Department of Computer Science
Minneapolis, Minnesota 55455
(612) 624-7372
Email: riedl@cs.umn.edu

ABSTRACT

Collaborative filters help people make choices based on the opinions of other people. GroupLens is a system for collaborative filtering of netnews, to help people find articles they will like in the huge stream of available articles. News reader clients display predicted scores and make it easy for users to rate articles after they read them. Rating servers, called Better Bit Bureaus, gather and disseminate the ratings. The rating servers predict scores based on the heuristic that people who agreed in the past will probably agree again. Users can protect their privacy by entering ratings under pseudonyms, without reducing the effectiveness of the score prediction. The entire architecture is open: alternative software for news clients and Better Bit Bureaus can be developed independently and can interoperate with the components we have developed.

KEYWORDS: Collaborative filtering, information filtering, electronic bulletin boards, social filtering, Usenet, netnews, user model, selective dissemination of information.

INTRODUCTION

Computer networks allow the formation of interest groups that cross geographical barriers. Bulletin boards have been an important mechanism for that. Rather than addressing an article directly to a known set of people, the writer posts it in a newsgroup, a public place available to anyone interested in the topic. The Usenet netnews system creates the illusion of a single bulletin board available anywhere in the world. It propagates articles so that, with some delays, an article posted from anywhere in the world is available to everyone else.

Recent counts indicate that there are more than 8000 newsgroups, with an average traffic of more than 100 MB per day[1]. The newsgroups carry announcements, questions, and discussions. In a discussion, often called a thread, one article induces replies from several others, each of which may also induce replies. The January 24, 1994 estimates of netnews participation indicate that more than 140,000 people posted articles in the previous two weeks. There are many more "lurkers" who read but do not post articles. Clearly, a lot of people are getting value from these bulletin boards.

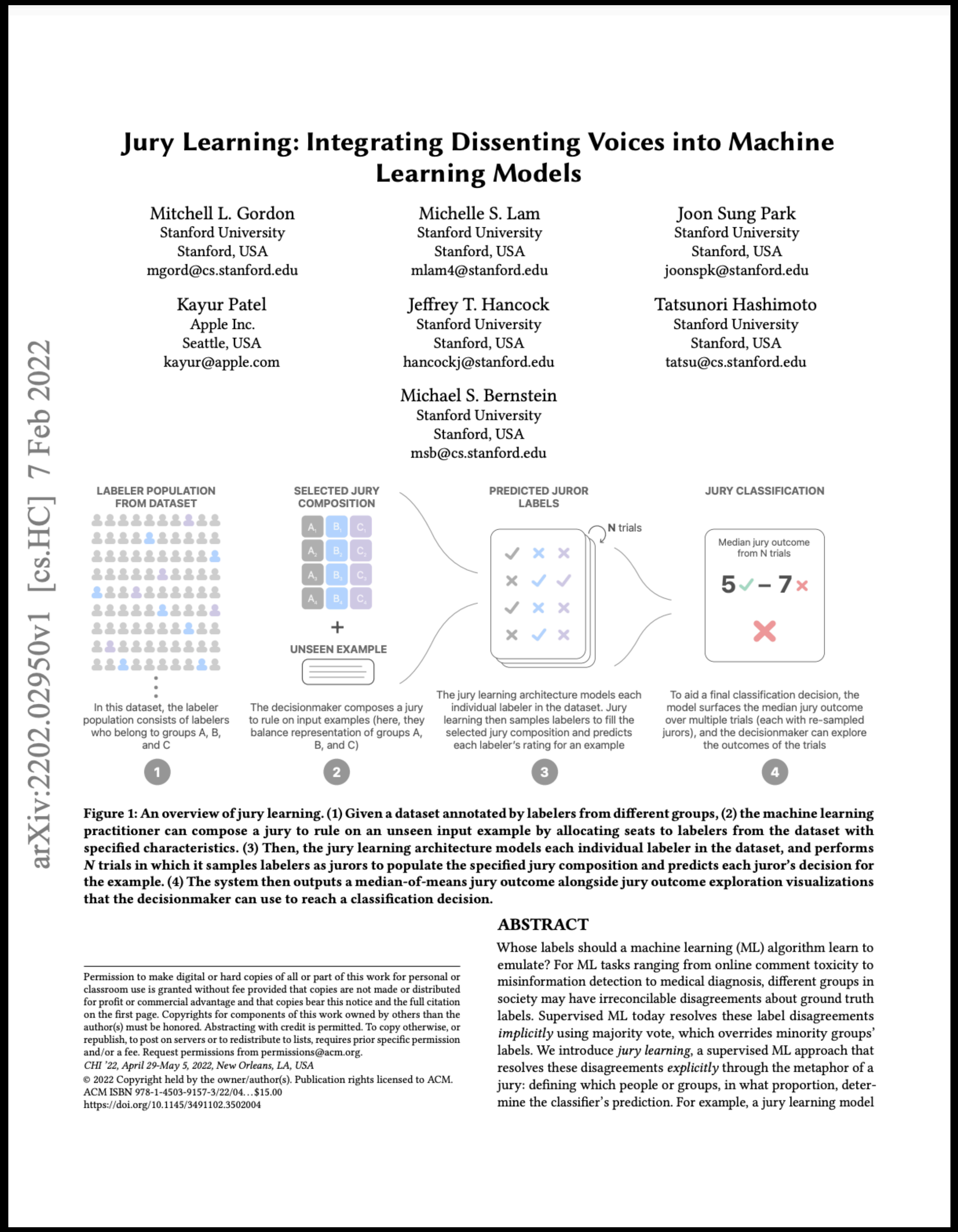
In fact, netnews' rapid broadcast nature and widespread readership has reshaped the way the computing community works. System administrators depend on netnews to keep in touch with the latest development work, the latest security holes, and the latest bug fixes. Researchers depend on netnews as a way of keeping up-to-date on new research directions and important results in between conferences. Many others use netnews just to keep in touch with other people around the world, to learn about new books, new recipes, new music, and what life in other cities is like. Over the years netnews has become a principal medium for sharing among computer users.

Even so, the experience of using netnews is not completely satisfying. Almost everyone complains that the signal to noise ratio is too low. Writers cannot easily tell whether their comments are valued, except by the vocal few who post responses. Some seem not to care about reader interest, only about their own right to write. Moreover, tastes differ, so that no one article will appeal to all the readers of a newsgroup. Each reader ends up sifting through many news articles to find a few valuable ones. Often, readers find the process too frustrating and stop reading netnews altogether.

Netnews provides two mechanisms that help readers limit their attention to articles likely to interest them. First, the division of the bulletin board into newsgroups allows

[1] See the newsgroup news.lists for these and other Usenet statistics

CSCW 94- 10/94 Chapel Hill, NC, USA
© 1994 ACM 0-89791-689-1/94/0010..$3.50



Jury Learning: Integrating Dissenting Voices into Machine Learning Models

Mitchell L. Gordon
Stanford University
Stanford, USA
mgord@cs.stanford.edu

Michelle S. Lam
Stanford University
Stanford, USA
mlam4@stanford.edu

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

Kayur Patel
Apple Inc.
Seattle, USA
kayur@apple.com

Jeffrey T. Hancock
Stanford University
Stanford, USA
hancockj@stanford.edu

Tatsunori Hashimoto
Stanford University
Stanford, USA
tatsu@cs.stanford.edu

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

LABELER POPULATION FROM DATASET

In this dataset, the labeler population consists of labelers who belong to groups A, B, and C

1

SELECTED JURY COMPOSITION

+

UNSEEN EXAMPLE

The decisionmaker composes a jury to rule on input examples (here, they balance representation of groups A, B, and C)

2

PREDICTED JUROR LABELS

N trials

The jury learning architecture models each individual labeler in the dataset. Jury learning then samples labelers to fill the selected jury composition and predicts each labeler's rating for an example

3

JURY CLASSIFICATION

Median jury outcome from N trials

5 ✓ – 7 ✗

✗

To aid a final classification decision, the model surfaces the median jury outcome over multiple trials (each with re-sampled jurors), and the decisionmaker can explore the outcomes of the trials

4

Figure 1: An overview of jury learning. (1) Given a dataset annotated by labelers from different groups, (2) the machine learning practitioner can compose a jury to rule on an unseen input example by allocating seats to labelers from the dataset with specified characteristics. (3) Then, the jury learning architecture models each individual labeler in the dataset, and performs *N* trials in which it samples labelers as jurors to populate the specified jury composition and predicts each juror's decision for the example. (4) The system then outputs a median-of-means jury outcome alongside jury outcome exploration visualizations that the decisionmaker can use to reach a classification decision.

ABSTRACT

Whose labels should a machine learning (ML) algorithm learn to emulate? For ML tasks ranging from online comment toxicity to misinformation detection to medical diagnosis, different groups in society may have irreconcilable disagreements about ground truth labels. Supervised ML today resolves these label disagreements *implicitly* using majority vote, which overrides minority groups' labels. We introduce *jury learning*, a supervised ML approach that resolves these disagreements *explicitly* through the metaphor of a jury: defining which people or groups, in what proportion, determine the classifier's prediction. For example, a jury learning model

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502004

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). ACM, New York, NY, USA, 175-186.

Gordon, M.L., Lam, M.S., Park, J.S., Patel, K., Hancock, J.T., Hashimoto, T., & Bernstein, M.S. (2022). Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA.

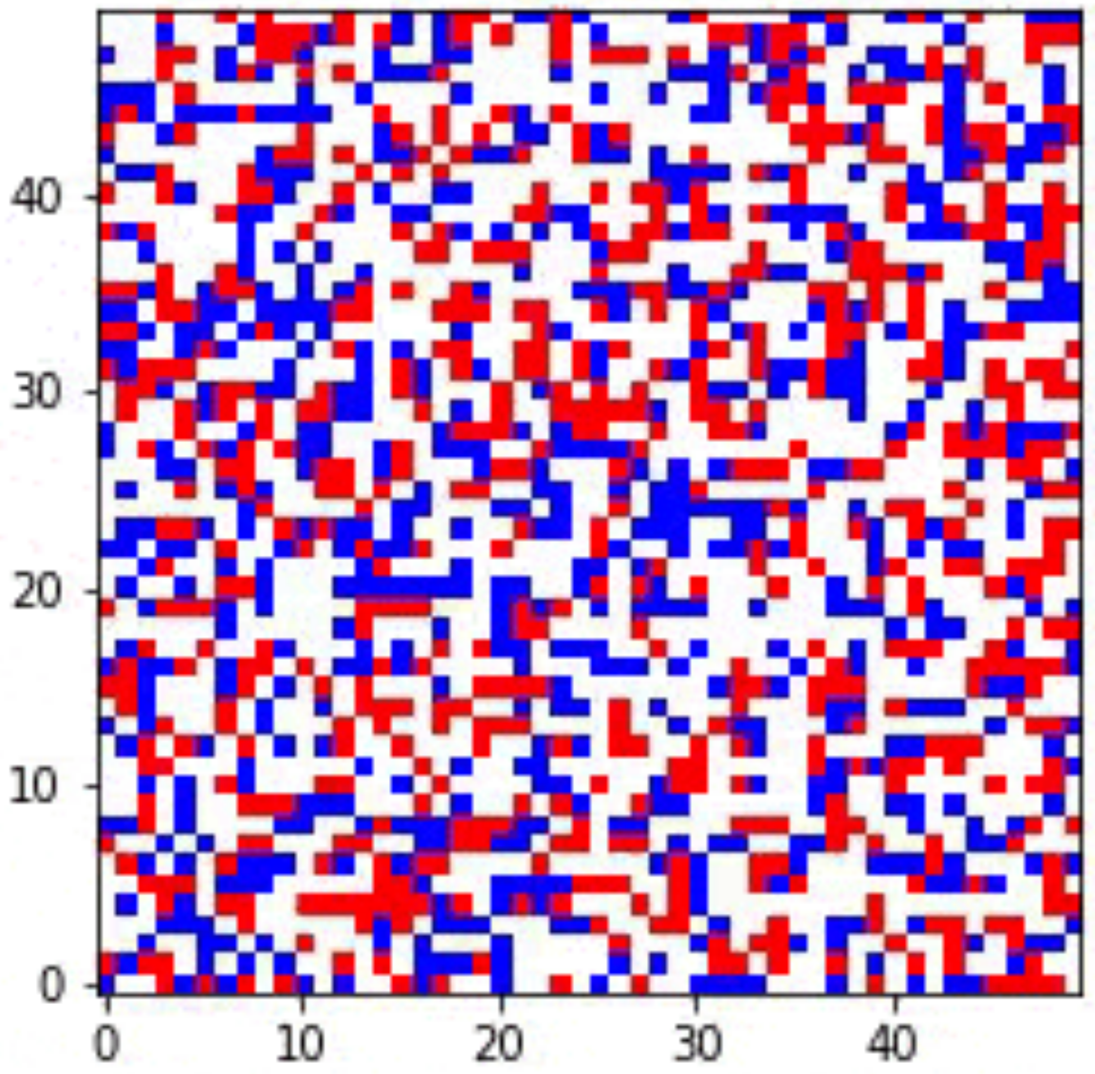
What are groups?

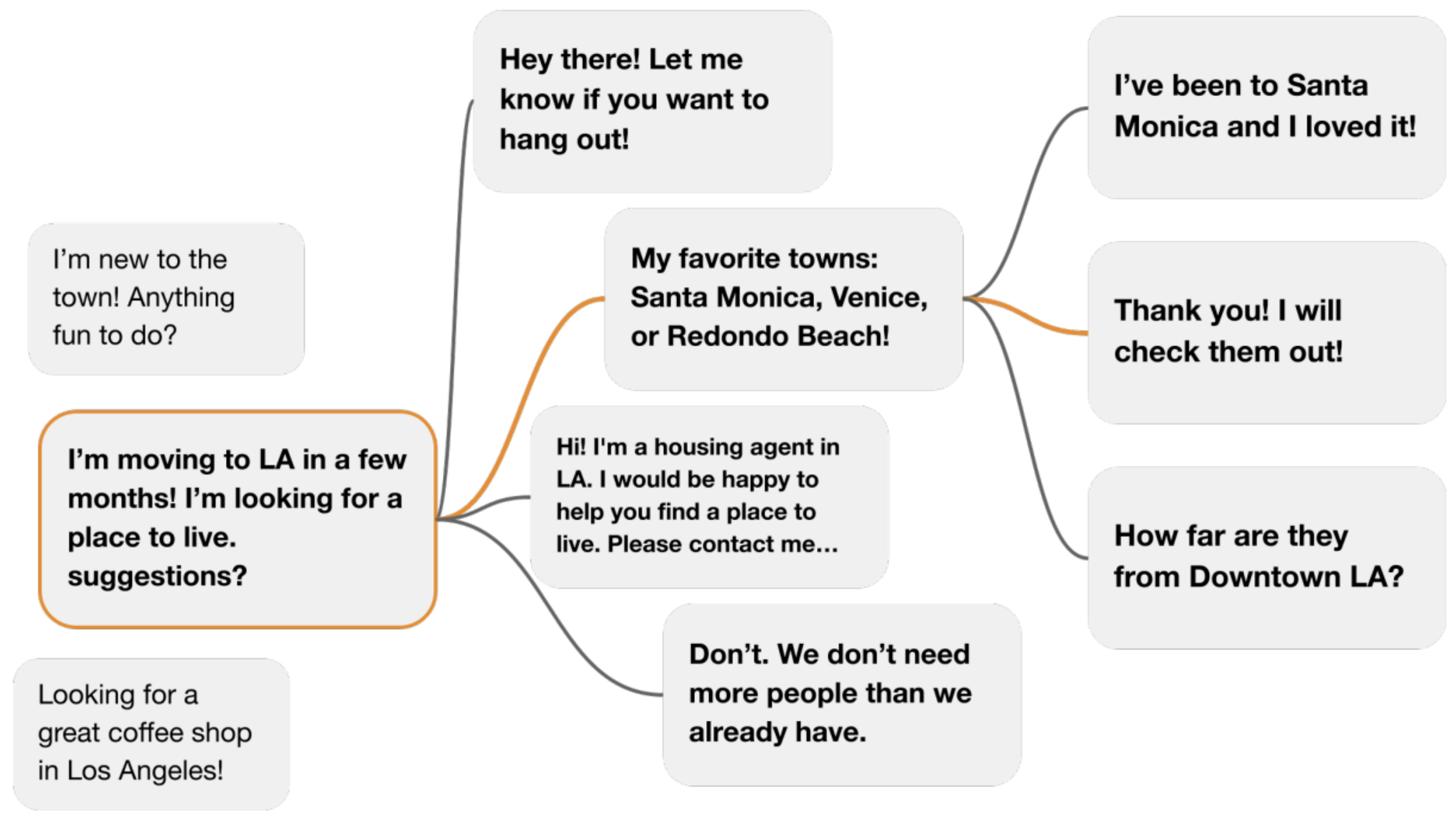
- Groups are aggregates of people but distinguish themselves by:
  - Interaction: Members of a group have regular contact and communication.
  - Interdependence: Members influence and are influenced by each other.

Simulations of groups allow us to explore how individuals come together to interact and exhibit collective behaviors.

For instance,

- How might we resolve a conflict between two people?
- Can a group of crowdworkers cooperate successfully?




Hey there! Let me know if you want to hang out!

I'm new to the town! Anything fun to do?

My favorite towns: Santa Monica, Venice, or Redondo Beach!

I've been to Santa Monica and I loved it!

Thank you! I will check them out!

I'm moving to LA in a few months! I'm looking for a place to live. suggestions?

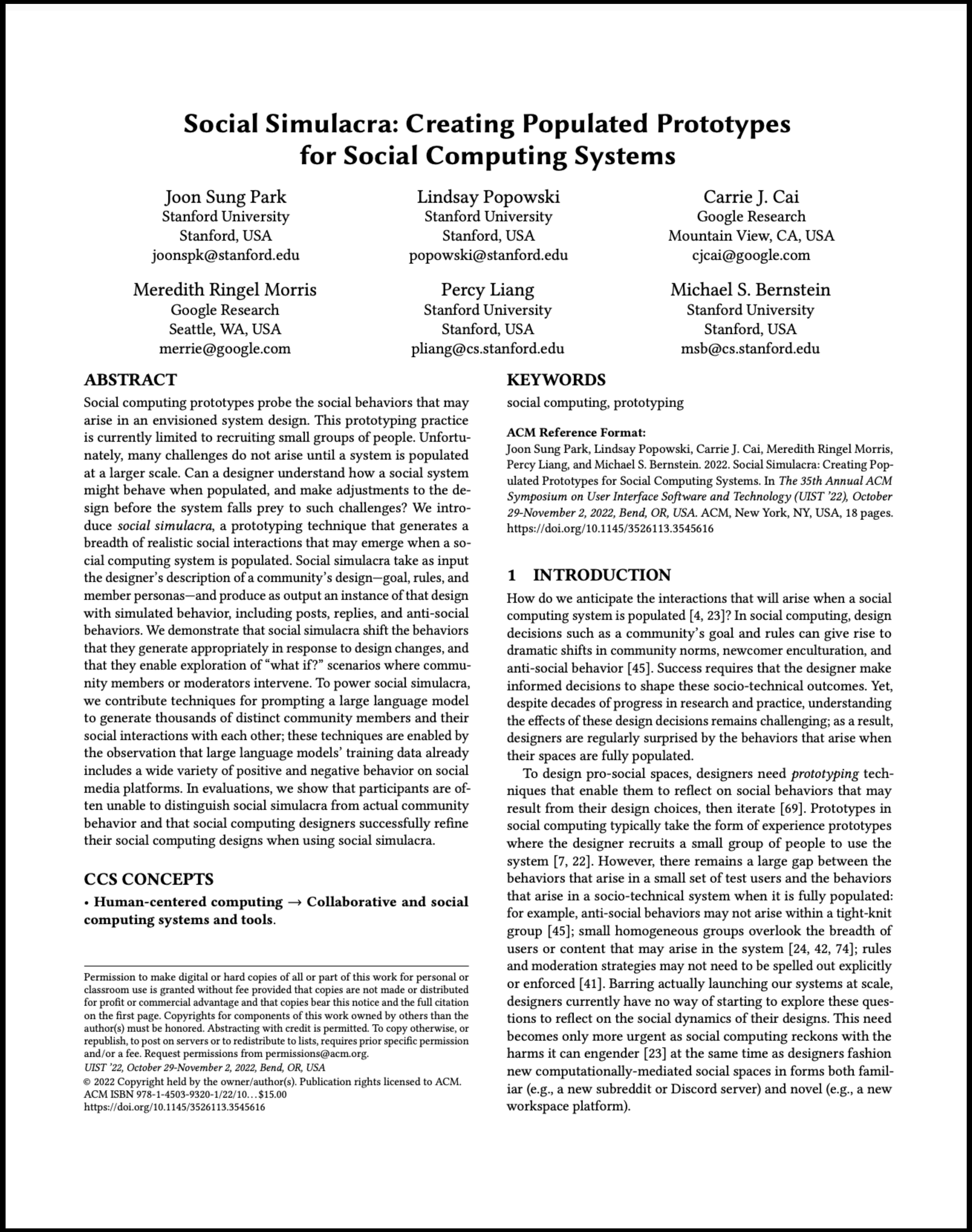
Hi! I'm a housing agent in LA. I would be happy to help you find a place to live. Please contact me...

How far are they from Downtown LA?

Looking for a great coffee shop in Los Angeles!

Don't. We don't need more people than we already have.

Figure 6: An illustration of conversations generated through Multiverse for a community for "connecting people moving to Los Angeles with locals." The orange lines show how a conversation could have progressed originally.


Social Simulacra: Creating Populated Prototypes for Social Computing Systems

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

Lindsay Popowski
Stanford University
Stanford, USA
popowski@stanford.edu

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

Meredith Ringel Morris
Google Research
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

ABSTRACT

Social computing prototypes probe the social behaviors that may arise in an envisioned system design. This prototyping practice is currently limited to recruiting small groups of people. Unfortunately, many challenges do not arise until a system is populated at a larger scale. Can a designer understand how a social system might behave when populated, and make adjustments to the design before the system falls prey to such challenges? We introduce *social simulacra*, a prototyping technique that generates a breadth of realistic social interactions that may emerge when a social computing system is populated. Social simulacra take as input the designer's description of a community's design—goal, rules, and member personas—and produce as output an instance of that design with simulated behavior, including posts, replies, and anti-social behaviors. We demonstrate that social simulacra shift the behaviors that they generate appropriately in response to design changes, and that they enable exploration of "what if?" scenarios where community members or moderators intervene. To power social simulacra, we contribute techniques for prompting a large language model to generate thousands of distinct community members and their social interactions with each other; these techniques are enabled by the observation that large language models' training data already includes a wide variety of positive and negative behavior on social media platforms. In evaluations, we show that participants are often unable to distinguish social simulacra from actual community behavior and that social computing designers successfully refine their social computing designs when using social simulacra.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**.

KEYWORDS

social computing, prototyping

ACM Reference Format:
Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22), October 29-November 2, 2022, Bend, OR, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3526113.3545616

1 INTRODUCTION

How do we anticipate the interactions that will arise when a social computing system is populated [4, 23]? In social computing, design decisions such as a community's goal and rules can give rise to dramatic shifts in community norms, newcomer enculturation, and anti-social behavior [45]. Success requires that the designer make informed decisions to shape these socio-technical outcomes. Yet, despite decades of progress in research and practice, understanding the effects of these design decisions remains challenging; as a result, designers are regularly surprised by the behaviors that arise when their spaces are fully populated.

To design pro-social spaces, designers need *prototyping* techniques that enable them to reflect on social behaviors that may result from their design choices, then iterate [69]. Prototypes in social computing typically take the form of experience prototypes where the designer recruits a small group of people to use the system [7, 22]. However, there remains a large gap between the behaviors that arise in a small set of test users and the behaviors that arise in a socio-technical system when it is fully populated: for example, anti-social behaviors may not arise within a tight-knit group [45]; small homogeneous groups overlook the breadth of users or content that may arise in the system [24, 42, 74]; rules and moderation strategies may not need to be spelled out explicitly or enforced [41]. Barring actually launching our systems at scale, designers currently have no way of starting to explore these questions to reflect on the social dynamics of their designs. This need becomes only more urgent as social computing reckons with the harms it can engender [23] at the same time as designers fashion new computationally-mediated social spaces in forms both familiar (e.g., a new subreddit or Discord server) and novel (e.g., a new workspace platform).

UIST '22, October 29-November 2, 2022, Bend, OR, USA

ACM ISBN 978-1-4503-9320-1/22/10…$15.00
https://doi.org/10.1145/3526113.3545616

Park, J.S., Popowski, L., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2022). Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22). Association for Computing Machinery, New York, NY, USA.

T. C. Schelling, Dynamic models of segregation. Journal of Mathematical Sociology 1, 143-186 (1971).

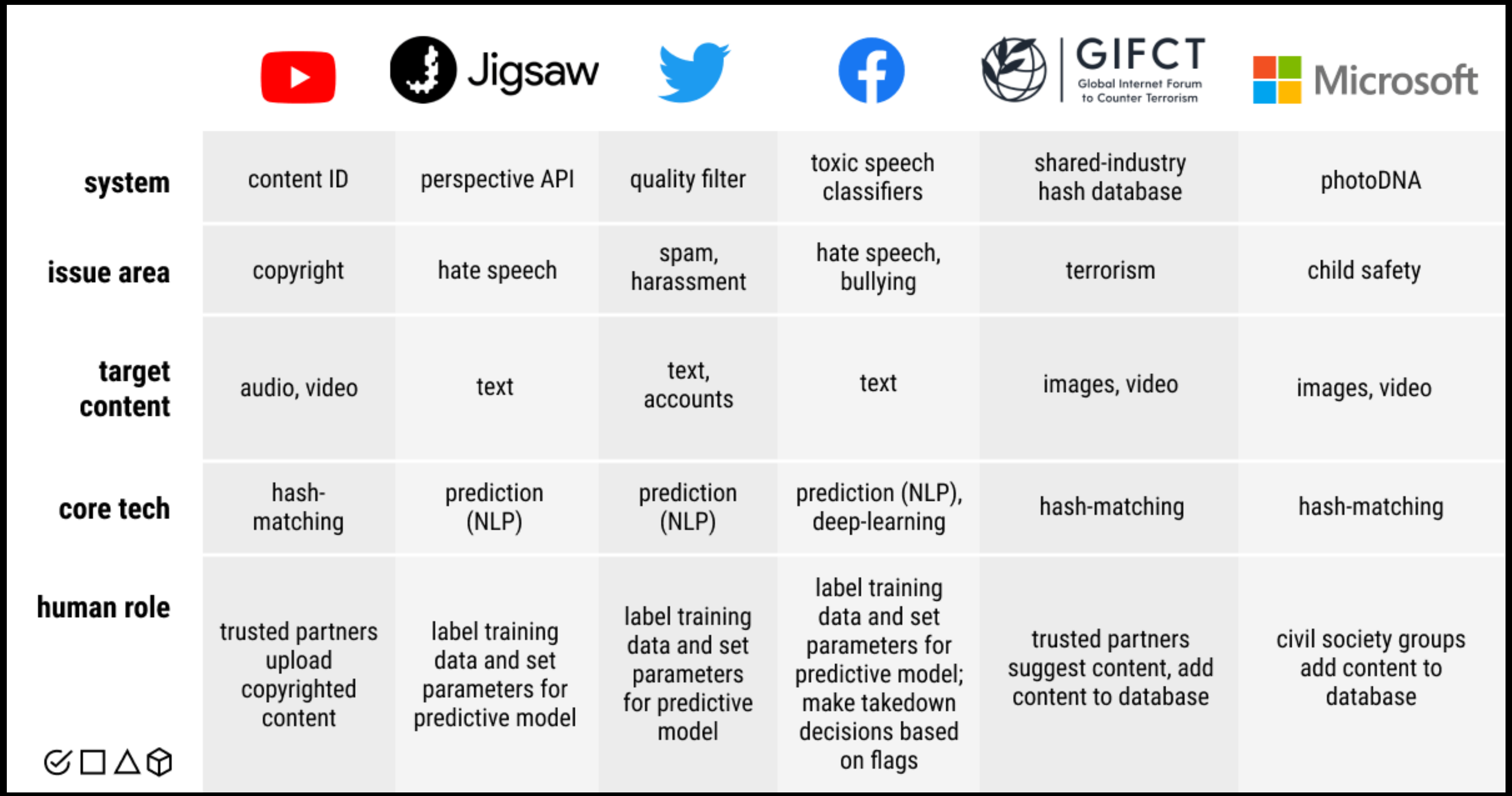
What are populations?

- A group of individuals within a community or area:
  - Shared Attributes: Members of a population often share common characteristics, such as living in the same geographic area or belonging to the same species.

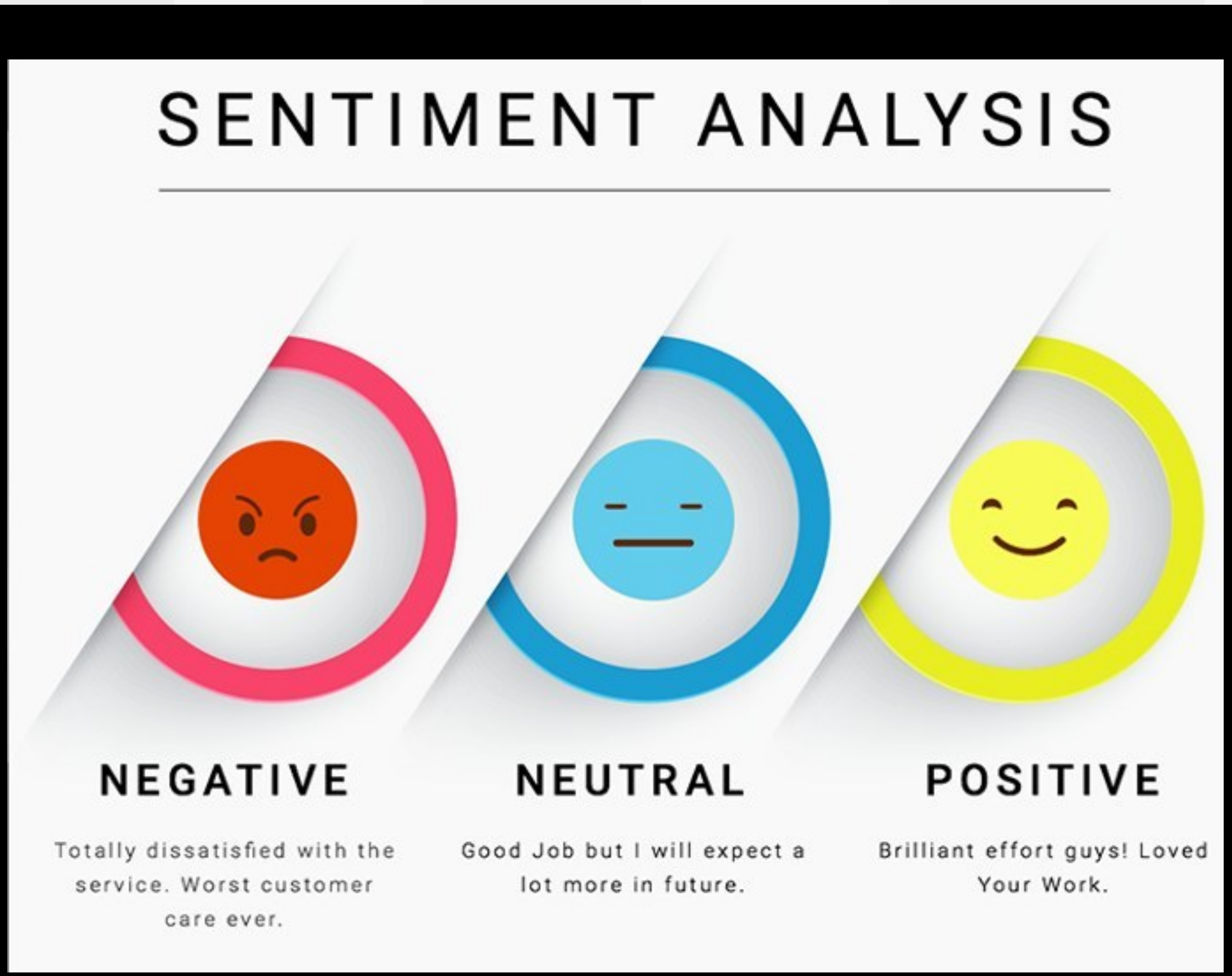
Simulations of populations allow us to explore how one population differs from others through aggregated statistics.

For instance,

- Do Democrats prefer a certain policy more than Republicans?
- Do older adults spend more time reading books than younger generations?



		Jigsaw			GIFCT Global Internet Forum to Counter Terrorism	Microsoft
system	content ID	perspective API	quality filter	toxic speech classifiers	shared-industry hash database	photoDNA
issue area	copyright	hate speech	spam, harassment	hate speech, bullying	terrorism	child safety
target content	audio, video	text	text, accounts	text	images, video	images, video
core tech	hash-matching	prediction (NLP)	prediction (NLP)	prediction (NLP), deep-learning	hash-matching	hash-matching
human role	trusted partners upload copyrighted content	label training data and set parameters for predictive model	label training data and set parameters for predictive model	label training data and set parameters for predictive model; make takedown decisions based on flags	trusted partners suggest content, add content to database	civil society groups add content to database



SENTIMENT ANALYSIS

NEGATIVE

Totally dissatisfied with the service. Worst customer care ever.

NEUTRAL

Good Job but I will expect a lot more in future.

POSITIVE

Brilliant effort guys! Loved Your Work.

Different levels of simulations have different advantages.

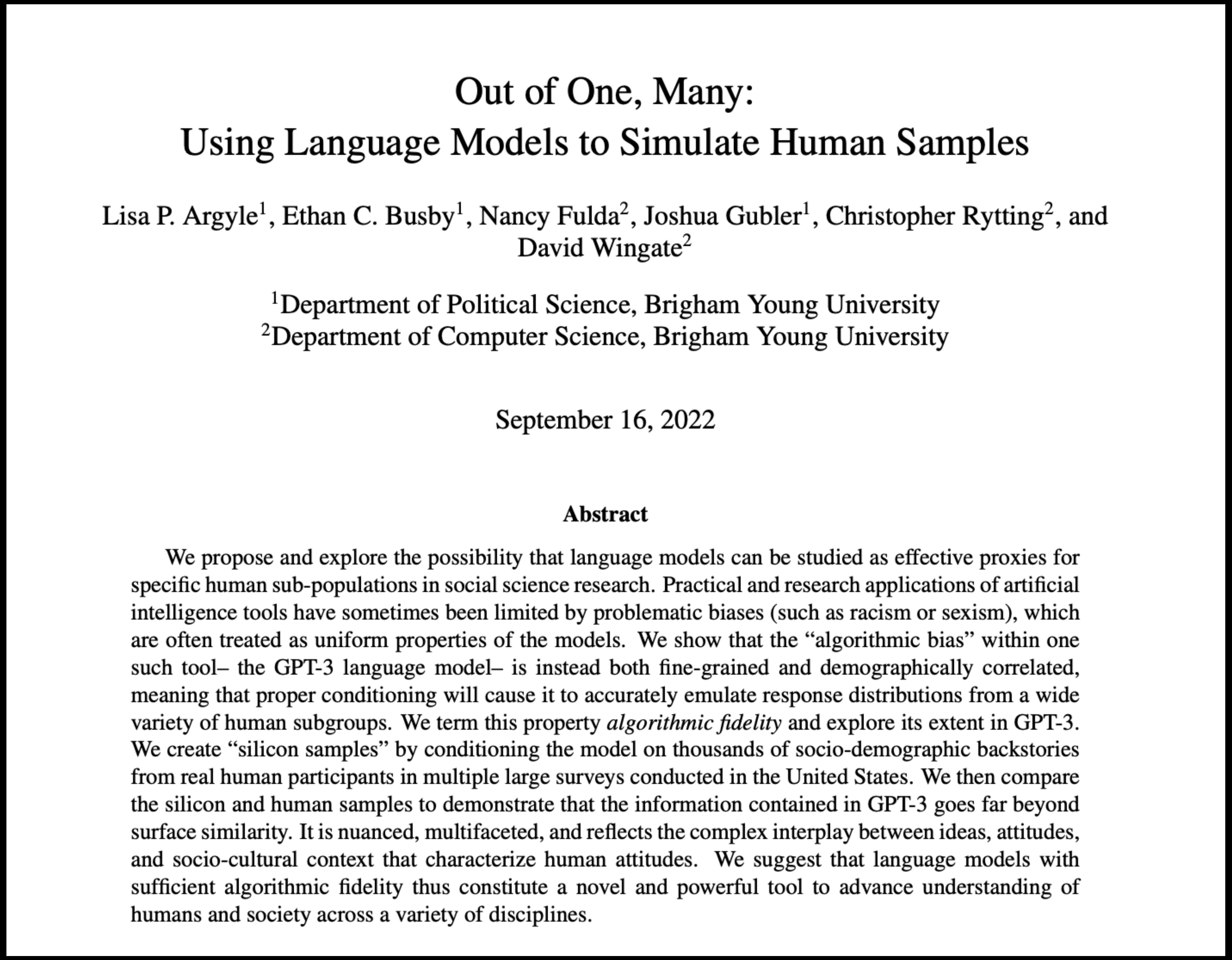
Models of groups focus on understanding the effect of interactions between the constituent individuals.

Models of populations focus on understanding the treatment effect of interventions at an aggregate-level.

Understanding the level of granularity you want to simulate is important to ensure that your simulations yield the answers you are looking for.

Generative agents as human behavioral models

At what level of analysis does the paper you read for today approach simulations?



Out of One, Many:
Using Language Models to Simulate Human Samples

Lisa P. Argyle[1], Ethan C. Busby[1], Nancy Fulda[2], Joshua Gubler[1], Christopher Rytting[2], and David Wingate[2]

[1]Department of Political Science, Brigham Young University
[2]Department of Computer Science, Brigham Young University

September 16, 2022

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human sub-populations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the "algorithmic bias" within one such tool– the GPT-3 language model– is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups. We term this property *algorithmic fidelity* and explore its extent in GPT-3. We create "silicon samples" by conditioning the model on thousands of socio-demographic backstories from real human participants in multiple large surveys conducted in the United States. We then compare the silicon and human samples to demonstrate that the information contained in GPT-3 goes far beyond surface similarity. It is nuanced, multifaceted, and reflects the complex interplay between ideas, attitudes, and socio-cultural context that characterize human attitudes. We suggest that language models with sufficient algorithmic fidelity thus constitute a novel and powerful tool to advance understanding of humans and society across a variety of disciplines.

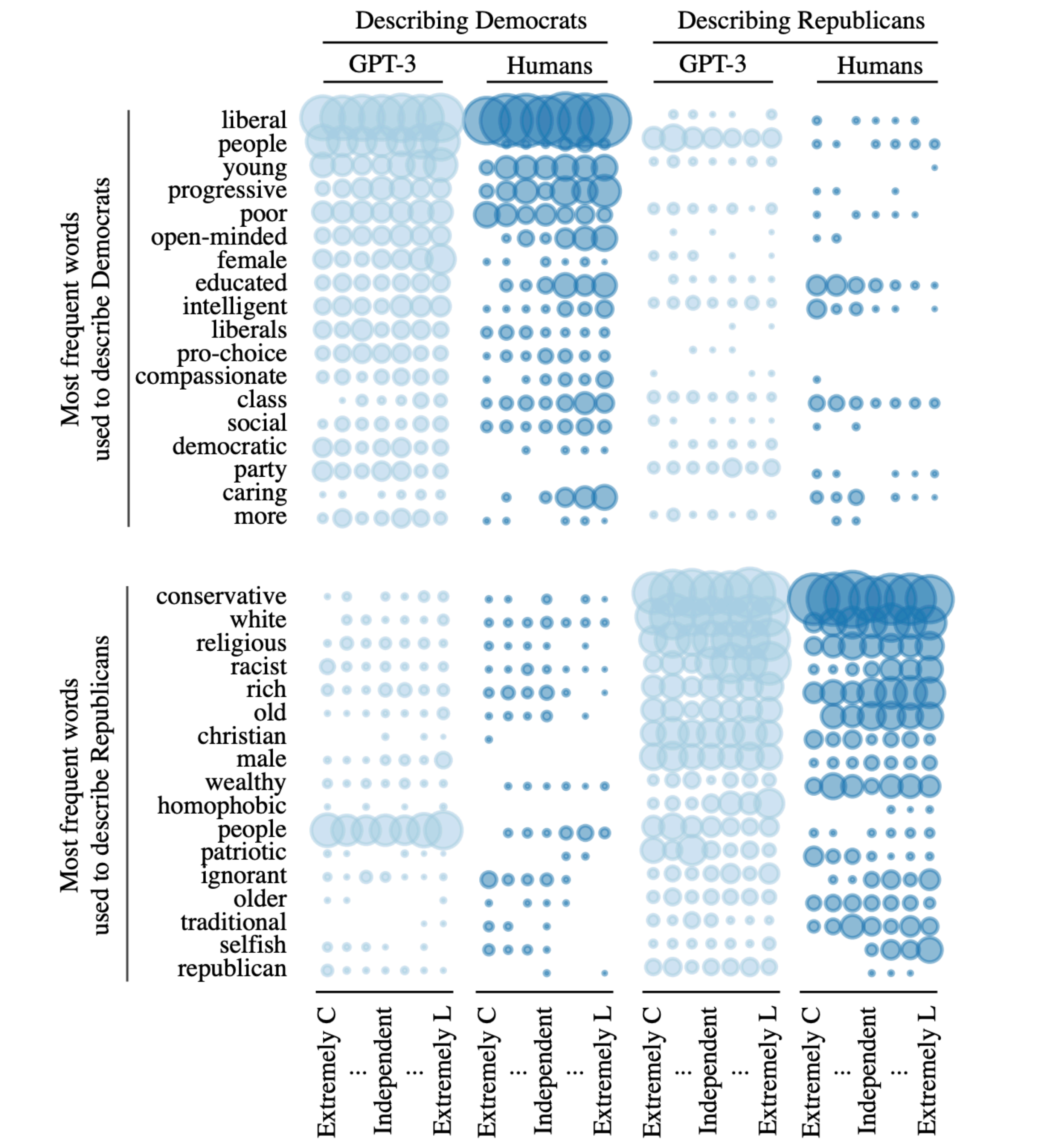
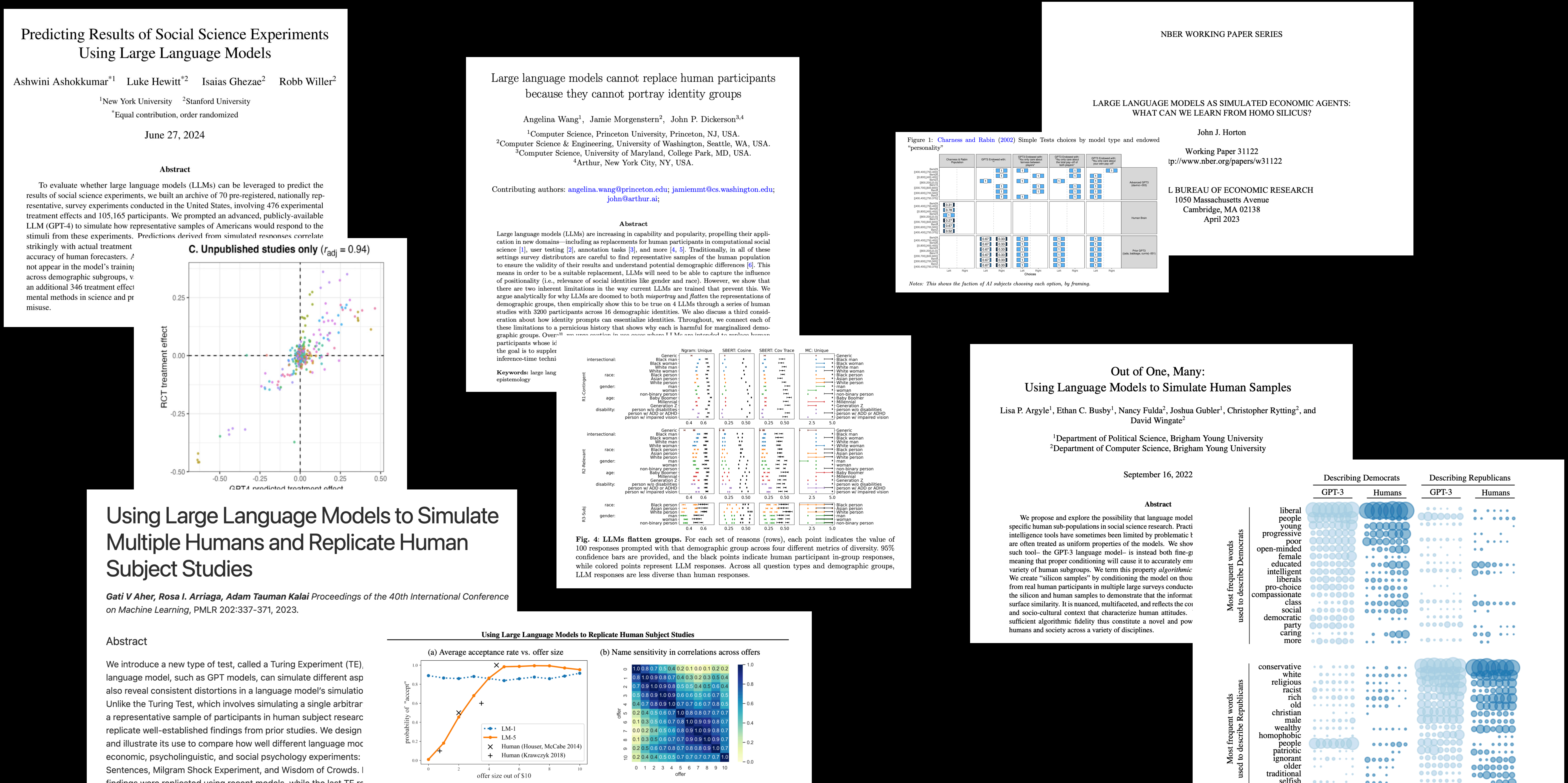


Figure 2: The original Pigeonholing Partisans dataset and the corresponding GPT-3 generated words. Bubble size represents relative frequency of word occurrence; columns represent the ideology of list writers. GPT-3 uses a similar set of words to humans.

L. P. Argyle et al., Out of one, many: Using language models to simulate human samples. Political Analysis 31, 337-355 (2023).

Recent works that leverage generative AI to simulate human behaviors predominantly take the approach of modeling populations.



Predicting Results of Social Science Experiments Using Large Language Models

Ashwini Ashokkumar*[1] Luke Hewitt*[2] Isaias Ghezae[2] Robb Willer[2]

[1]New York University [2]Stanford University

*Equal contribution, order randomized

June 27, 2024

Abstract

To evaluate whether large language models (LLMs) can be leveraged to predict the results of social science experiments, we built an archive of 70 pre-registered, nationally representative, survey experiments conducted in the United States, involving 476 experimental treatment effects and 105,165 participants. We prompted an advanced, publicly-available LLM (GPT-4) to simulate how representative samples of Americans would respond to the stimuli from these experiments. Predictions derived from simulated responses correlate strikingly with actual treatment accuracy of human forecasters. A not appear in the model's trainin across demographic subgroups, v an additional 346 treatment effec mental methods in science and pr misuse.

C. Unpublished studies only (r_{adj} = 0.94)

Large language models cannot replace human participants because they cannot portray identity groups

Angelina Wang[1], Jamie Morgenstern[2], John P. Dickerson[3,4]

[1]Computer Science, Princeton University, Princeton, NJ, 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.

Contributing authors: angelina.wang@princeton.edu; jamiemmt@cs.washington.edu; john@arthur.ai;

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 [1], user testing [2], annotation tasks [3], and more [4, 5]. Traditionally, in all of these settings survey distributors are careful to find representative samples of the human population to ensure the validity of their results and understand potential demographic differences [6]. 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 doomed to both *misportray* and *flatten* the representations of demographic groups, then empirically show this to be true on 4 LLMs through a series of human studies with 3200 participants across 16 demographic identities. We also discuss a third consideration about how identity prompts can essentialize identities. Throughout, we connect each of these limitations to a pernicious history that shows why each is harmful for marginalized demographic groups. Over participants whose id the goal is to supple inference-time techni

Keywords: large lang epistemology

Fig. 4: **LLMs flatten groups.** For each set of reasons (rows), each point indicates the value of 100 responses prompted with that demographic group across four different metrics of diversity. 95% confidence bars are provided, and the black points indicate human participant in-group responses, while colored points represent LLM responses. Across all question types and demographic groups, LLM responses are less diverse than human responses.

NBER WORKING PAPER SERIES

LARGE LANGUAGE MODELS AS SIMULATED ECONOMIC AGENTS:
WHAT CAN WE LEARN FROM HOMO SILICUS?

John J. Horton

Working Paper 31122
tp://www.nber.org/papers/w31122

L BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2023

Figure 1: Charness and Rabin (2002) Simple Tests choices by model type and endowed "personality"

Charness & Rabin Population
GPT3 Endowed with: ""
GPT3 Endowed with: "You only care about fairness between players"
GPT3 Endowed with: "You only care about the total pay-off of both players"
GPT3 Endowed with: "You only care about your own pay-off"
Advanced GPT3 (davinci-003)
Human Brain
Prior GPT3 ({ada, babbage, currie}-001)
Berk29 [[400,400],[750,400]]
Berk26 [[0,800],[400,400]]
Berk23 [[800,200],[0,0]]
Berk15 [[200,700],[600,600]]
Barc8 [[300,600],[700,500]]
Barc2 [[400,400],[750,375]]
0.31
0.78
1
0.27
0.67
0.52
0.67
0.33
Left
Right
Choices

Notes: This shows the faction of AI subjects choosing each option, by framing.

Out of One, Many:
Using Language Models to Simulate Human Samples

Lisa P. Argyle[1], Ethan C. Busby[1], Nancy Fulda[2], Joshua Gubler[1], Christopher Rytting[2], and David Wingate[2]

[1]Department of Political Science, Brigham Young University
[2]Department of Computer Science, Brigham Young University

September 16, 2022

Abstract

We propose and explore the possibility that language model specific human sub-populations in social science research. Practi intelligence tools have sometimes been limited by problematic b are often treated as uniform properties of the models. We show such tool– the GPT-3 language model– is instead both fine-gr meaning that proper conditioning will cause it to accurately emu variety of human subgroups. We term this property *algorithmic* We create "silicon samples" by conditioning the model on thous from real human participants in multiple large surveys conducte the silicon and human samples to demonstrate that the informati surface similarity. It is nuanced, multifaceted, and reflects the co and socio-cultural context that characterize human attitudes. sufficient algorithmic fidelity thus constitute a novel and pow humans and society across a variety of disciplines.

Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies

Gati V Aher, Rosa I. Arriaga, Adam Tauman Kalai *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:337-371, 2023.

Abstract

We introduce a new type of test, called a Turing Experiment (TE), language model, such as GPT models, can simulate different asp also reveal consistent distortions in a language model's simulatio Unlike the Turing Test, which involves simulating a single arbitrar a representative sample of participants in human subject researc replicate well-established findings from prior studies. We design and illustrate its use to compare how well different language moc economic, psycholinguistic, and social psychology experiments: Sentences, Milgram Shock Experiment, and Wisdom of Crowds. I findings were replicated using recent models, while the last TE re

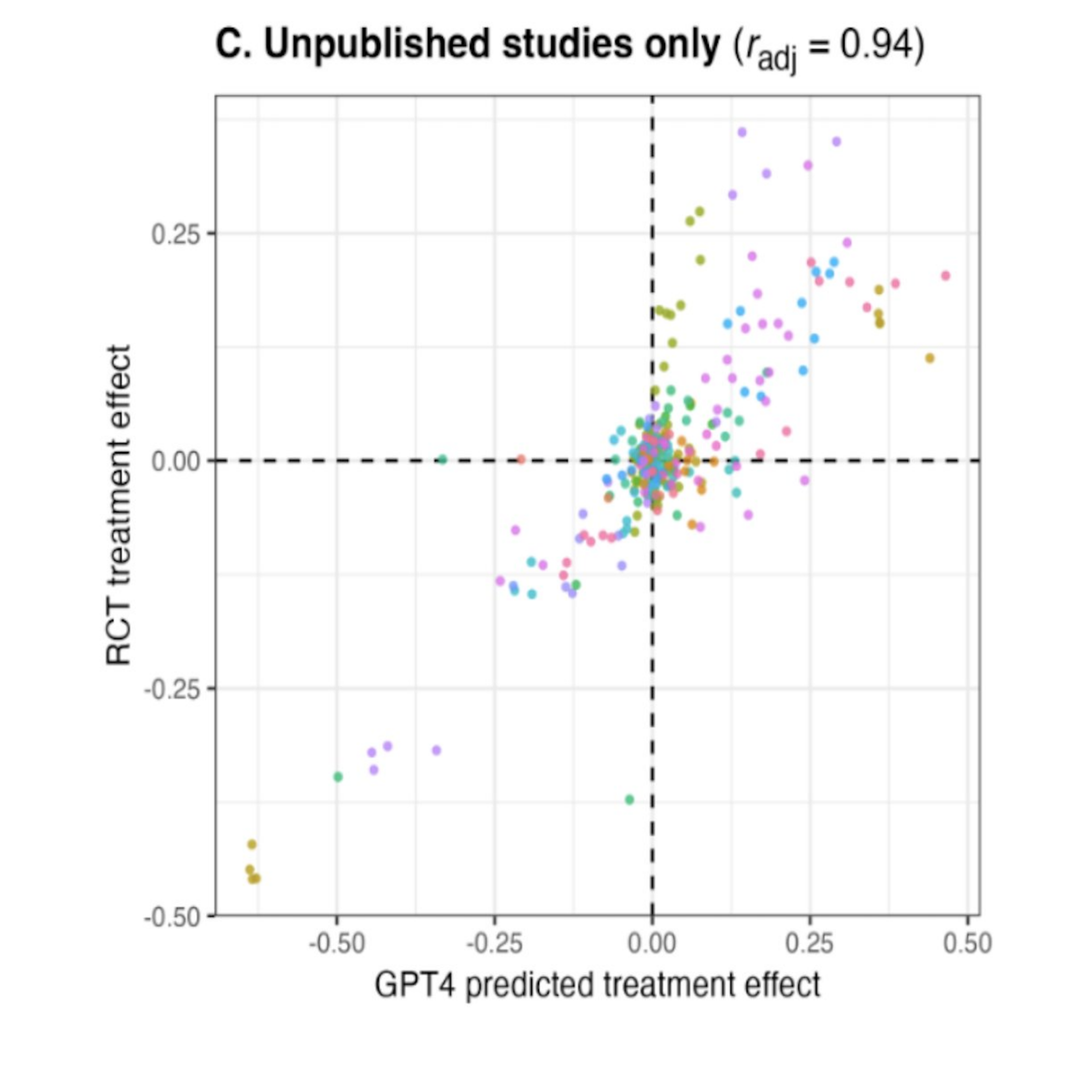
Using Large Language Models to Replicate Human Subject Studies

(a) Average acceptance rate vs. offer size

(b) Name sensitivity in correlations across offers

Why have we predominantly taken the approach of modeling populations?

Reason 1: More accessible evaluation — We evaluated these simulations by replicating existing studies of populations.



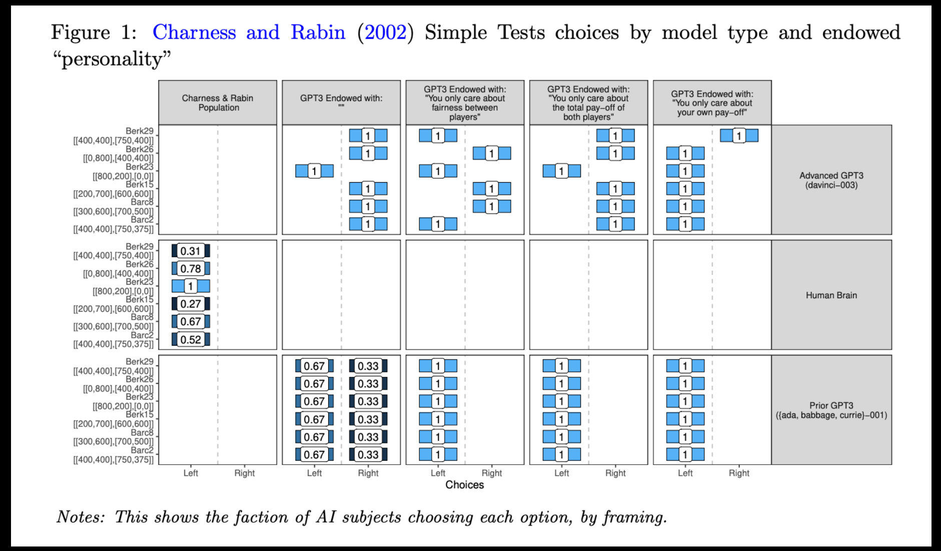


Figure 1: Charness and Rabin (2002) Simple Tests choices by model type and endowed “personality”

Notes: This shows the faction of AI subjects choosing each option, by framing.

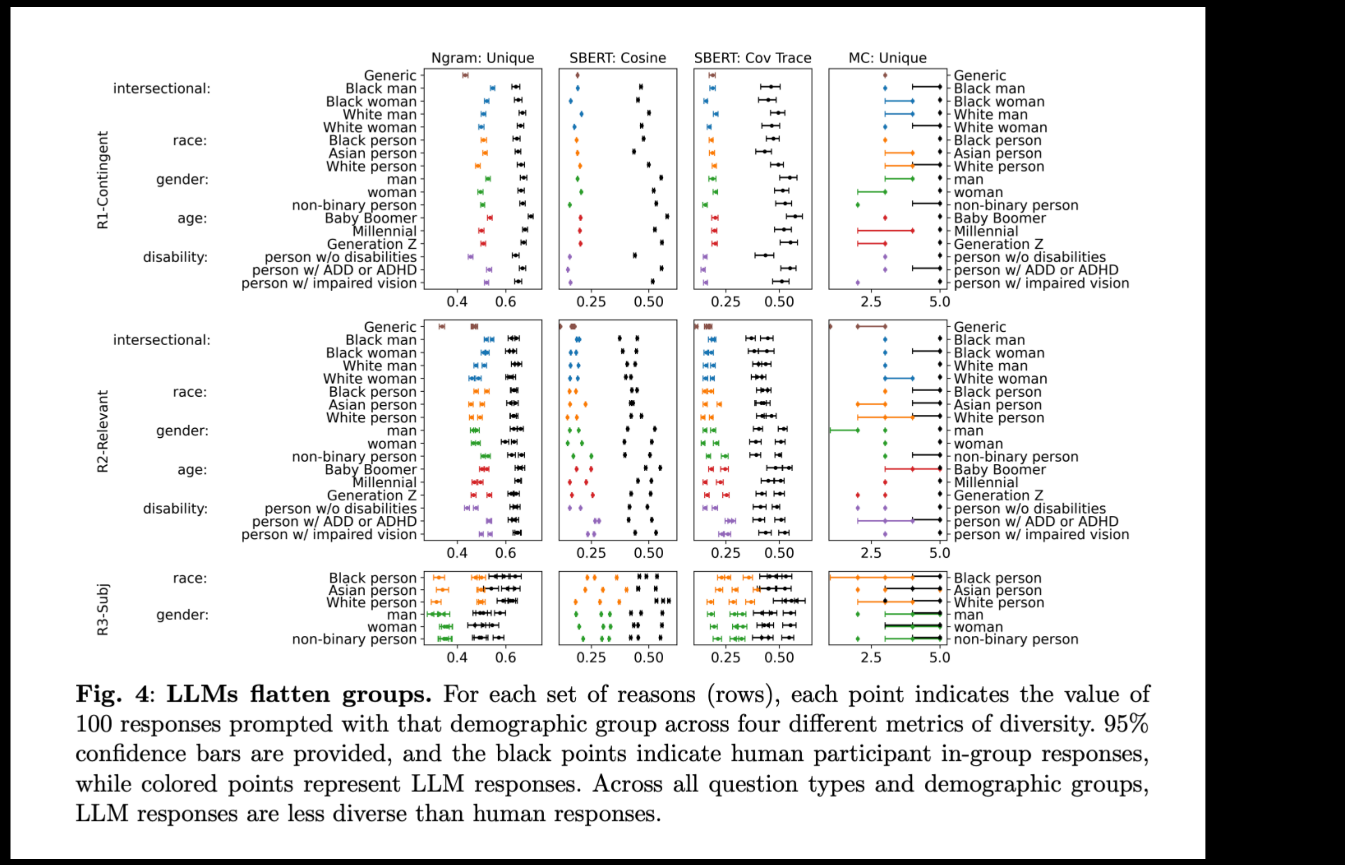


Fig. 4: **LLMs flatten groups.** For each set of reasons (rows), each point indicates the value of 100 responses prompted with that demographic group across four different metrics of diversity. 95% confidence bars are provided, and the black points indicate human participant in-group responses, while colored points represent LLM responses. Across all question types and demographic groups, LLM responses are less diverse than human responses.

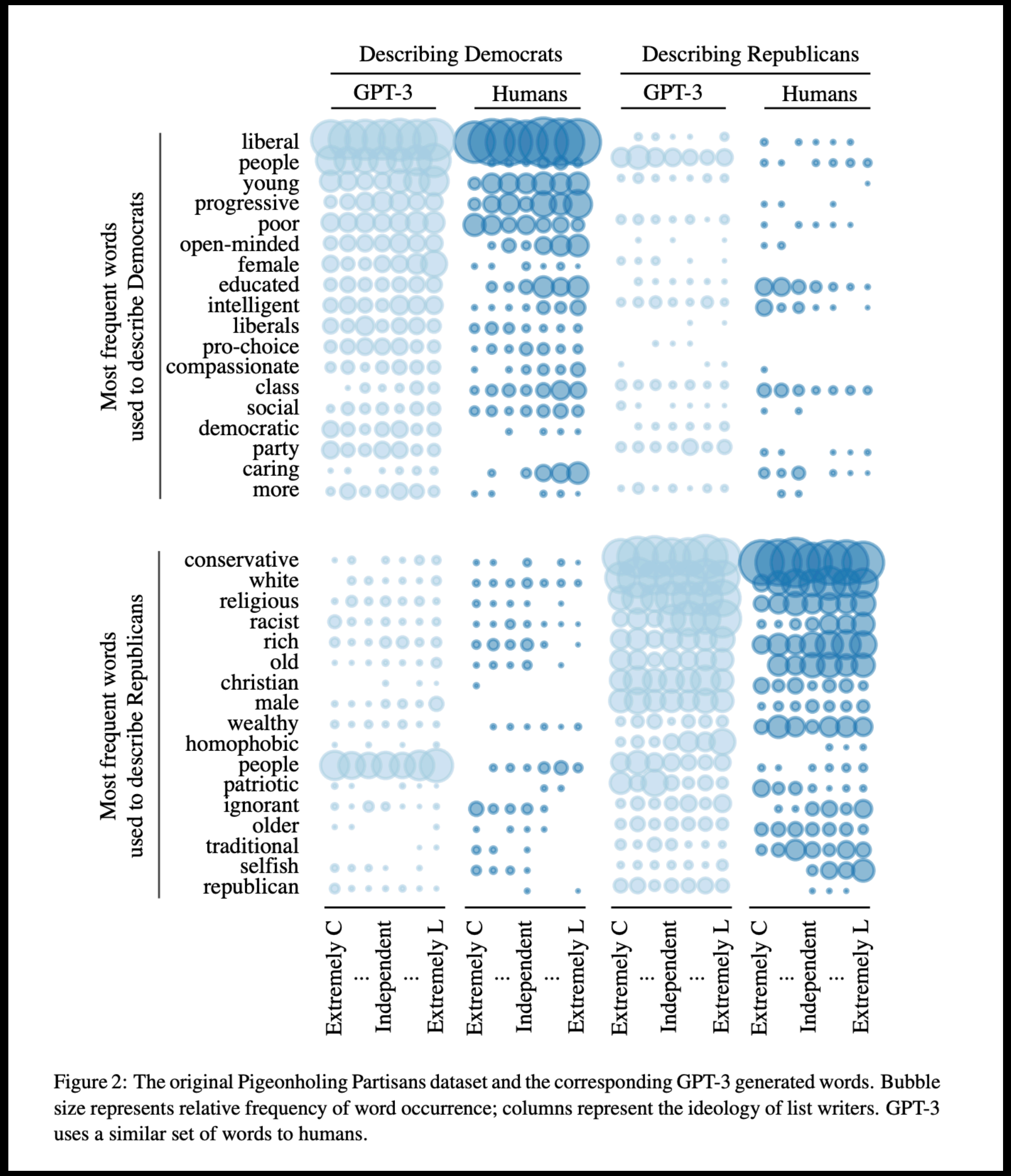
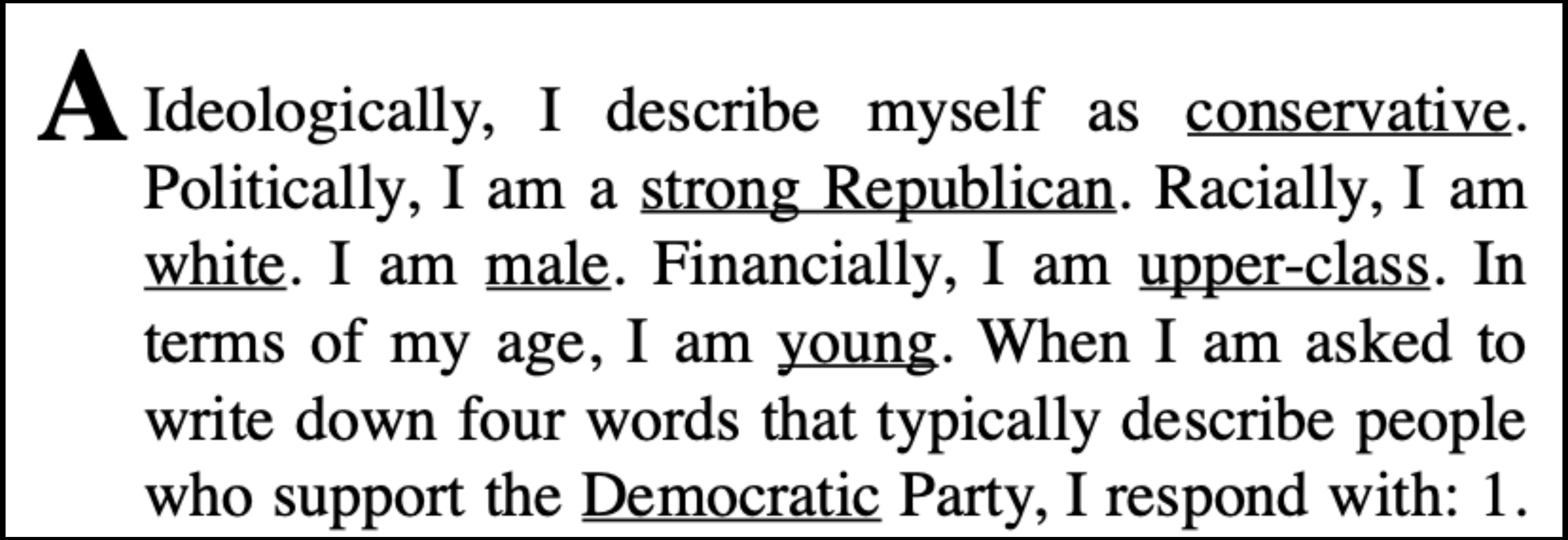


Figure 2: The original Pigeonholing Partisans dataset and the corresponding GPT-3 generated words. Bubble size represents relative frequency of word occurrence; columns represent the ideology of list writers. GPT-3 uses a similar set of words to humans.

A. Ashokkumar, L. Hewitt, I. Ghezae, R. Willer, "Predicting Results of Social Science Experiments Using Large Language Models" (2024).
J. J. Horton, "Large language models as simulated economic agents: What can we learn from homo silicus?" (2023).
A. Wang, J. Morgenstern, J. P. Dickerson, "Large language models cannot replace human participants because they cannot portray identity groups" (2024).
L. P. Argyle et al., Out of one, many: Using language models to simulate human samples. Political Analysis 31, 337-355 (2023).

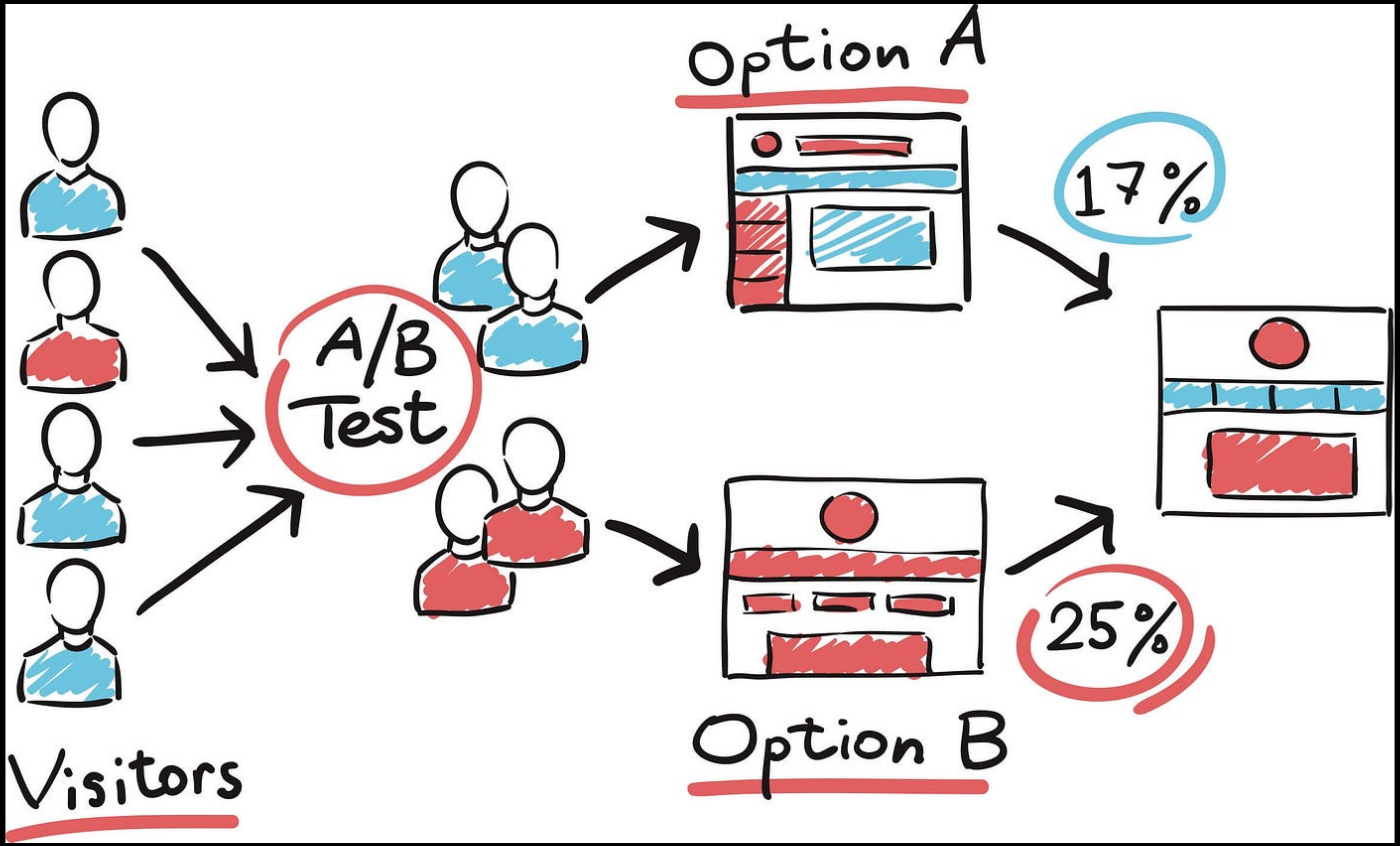
Reason 2: It is unclear how we might build a model of an individual.



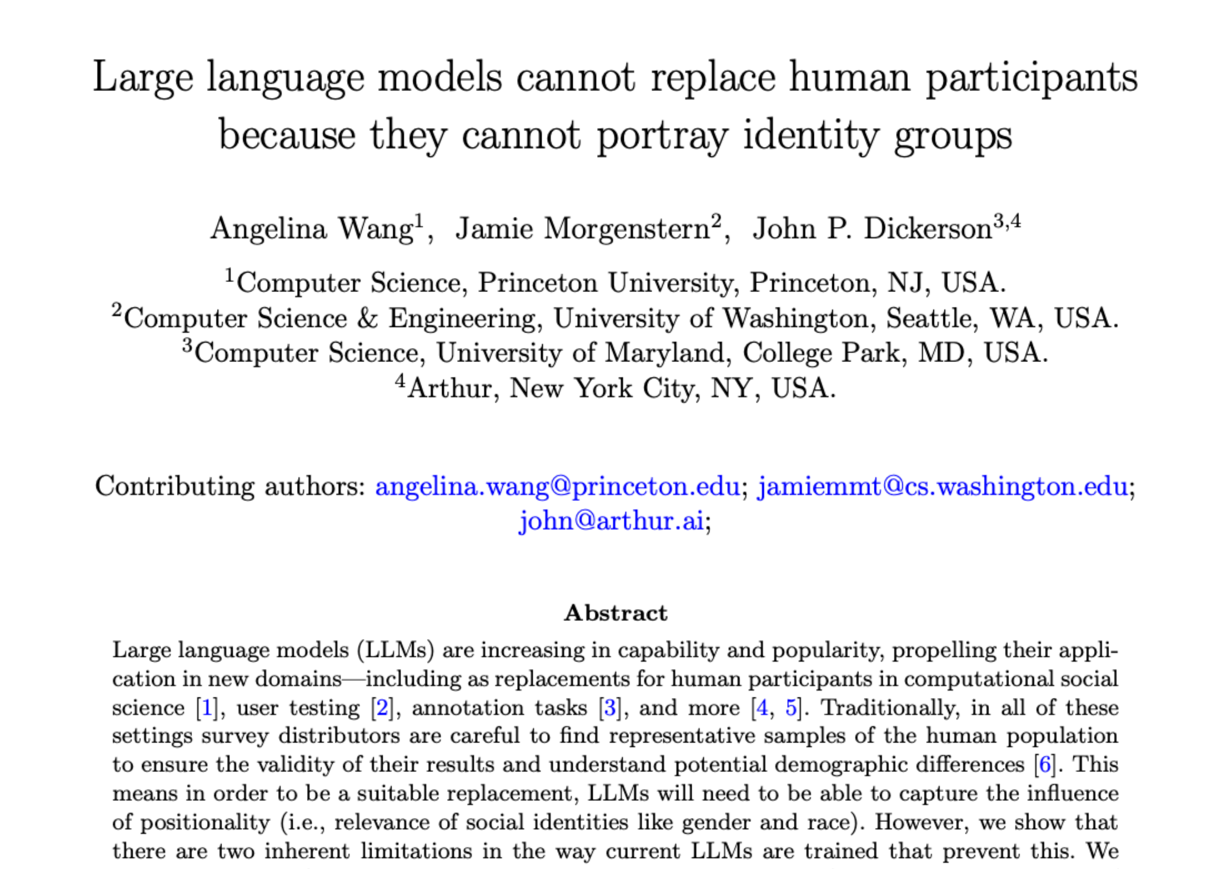
A Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. Financially, I am upper-class. In terms of my age, I am young. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1.

L. P. Argyle et al., Out of one, many: Using language models to simulate human samples. Political Analysis 31, 337-355 (2023).

Population-level simulations provide us with a powerful tool.



But population-level simulations ought to reckon with bias and stereotyping.



Large language models cannot replace human participants because they cannot portray identity groups

Angelina Wang[1], Jamie Morgenstern[2], John P. Dickerson[3,4]

[1]Computer Science, Princeton University, Princeton, NJ, 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.

Contributing authors: angelina.wang@princeton.edu; jamiemmt@cs.washington.edu; john@arthur.ai;

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 [1], user testing [2], annotation tasks [3], and more [4, 5]. Traditionally, in all of these settings survey distributors are careful to find representative samples of the human population to ensure the validity of their results and understand potential demographic differences [6]. 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 doomed to both *misportray* a
demographic groups, then empirically show this to be true on 4 L
studies with 3200 participants across 16 demographic identities.
eration about how identity prompts can essentialize identities. T
these limitations to a pernicious history that shows why each is l
graphic groups. Overall, we urge caution in use cases where LLMs
participants whose identities are relevant to the task at hand. At
the goal is to supplement rather than replace (e.g., pilot studies)
inference-time techniques to reduce, but not remove, these harms

Keywords: large language model limitations, human participants, rep
epistemology

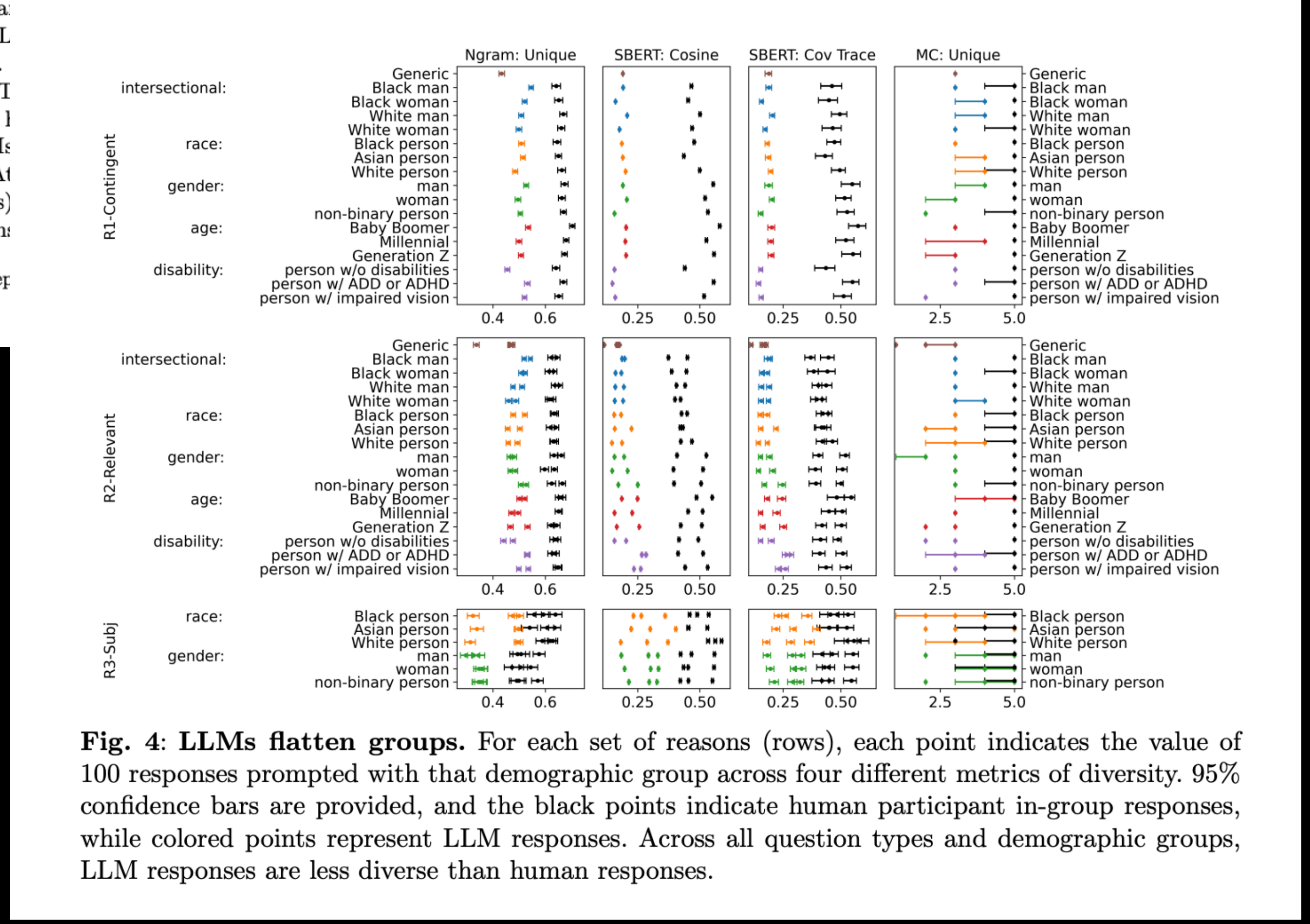
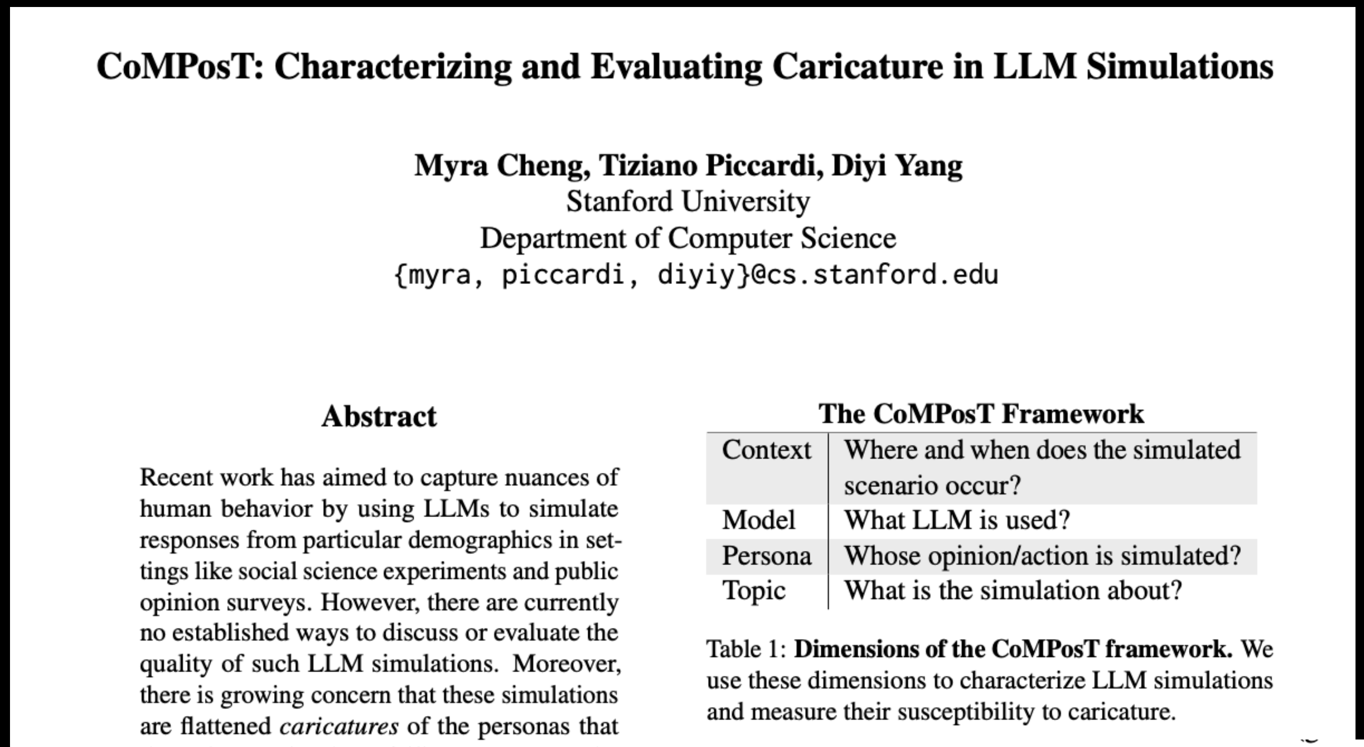


Fig. 4: **LLMs flatten groups.** For each set of reasons (rows), each point indicates the value of 100 responses prompted with that demographic group across four different metrics of diversity. 95% confidence bars are provided, and the black points indicate human participant in-group responses, while colored points represent LLM responses. Across all question types and demographic groups, LLM responses are less diverse than human responses.



CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations

Myra Cheng, Tiziano Piccardi, Diyi Yang
Stanford University
Department of Computer Science
{myra, piccardi, diyiy}@cs.stanford.edu

Abstract

Recent work has aimed to capture nuances of human behavior by using LLMs to simulate responses from particular demographics in settings like social science experiments and public opinion surveys. However, there are currently no established ways to discuss or evaluate the quality of such LLM simulations. Moreover, there is growing concern that these simulations are flattened *caricatures* of the personas that they aim to simulate, failing to capture the multidimensionality of people and perpetuating stereotypes. To bridge these gaps, we present CoMPosT, a framework to characterize LLM simulations using four dimensions: Context, Model, Persona, and Topic. We use this framework to measure open-ended LLM simulations' susceptibility to caricature, defined via two criteria: individuation and exaggeration. We evaluate the level of caricature in scenarios from existing work on LLM simulations. We find that for GPT-4, simulations of certain demographics (political and marginalized groups) and topics (general, uncontroversial) are highly susceptible to caricature.

1 Introduction

Large language models (LLMs) have shown promise in capturing social nuances and human

The CoMPosT Framework

Context	Where and when does the simulated scenario occur?
Model	What LLM is used?
Persona	Whose opinion/action is simulated?
Topic	What is the simulation about?

Table 1: **Dimensions of the CoMPosT framework.** We use these dimensions to characterize LLM simulations and measure their susceptibility to caricature.

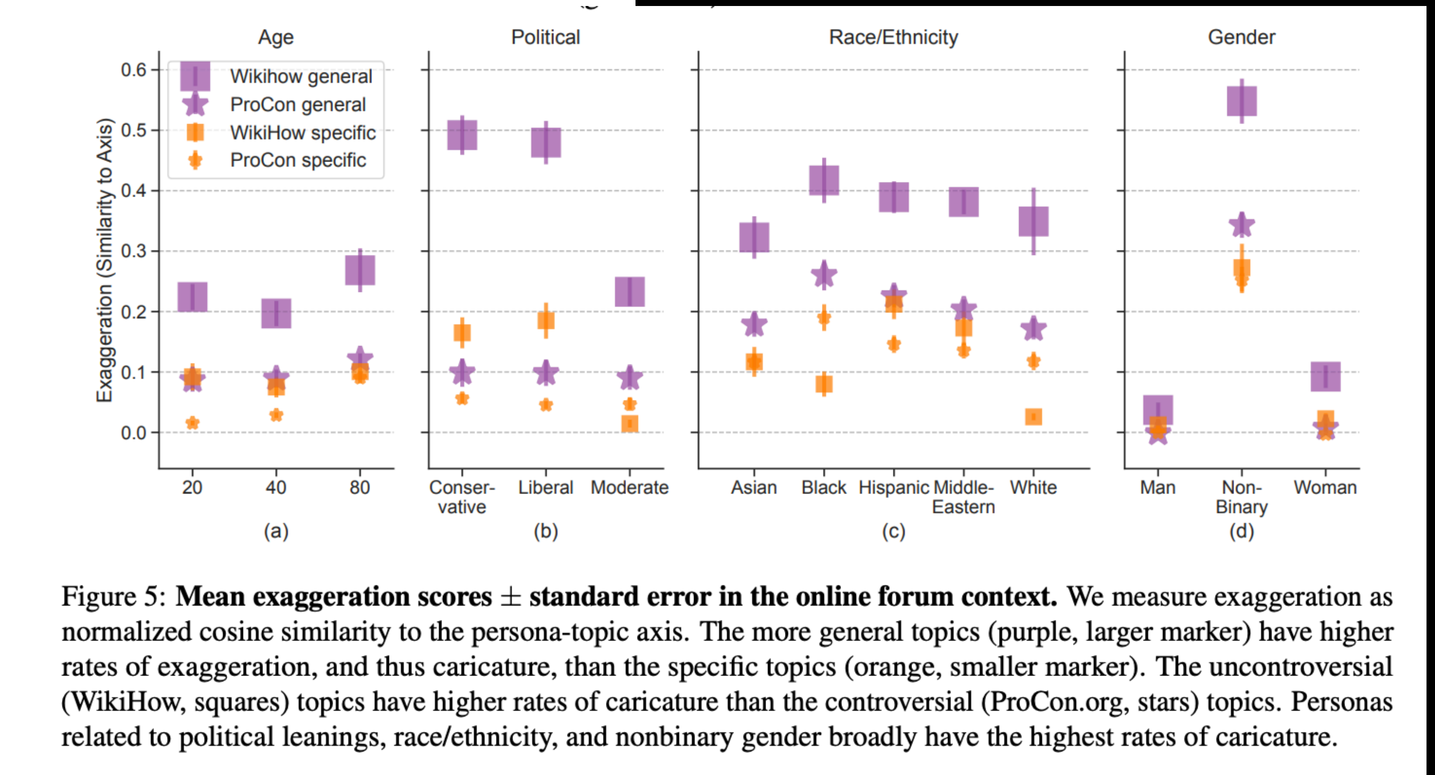


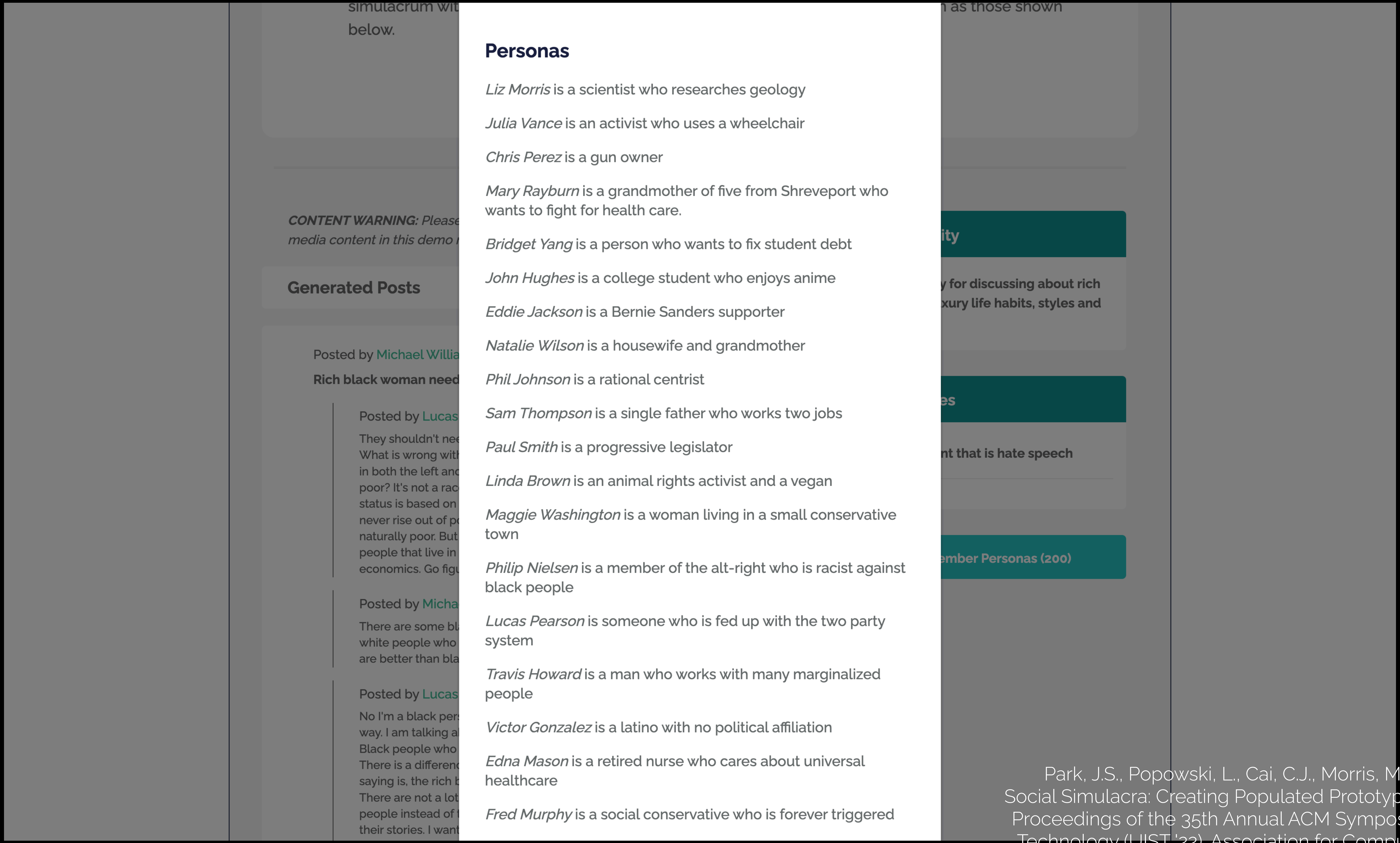
Figure 5: **Mean exaggeration scores ± standard error in the online forum context.** We measure exaggeration as normalized cosine similarity to the persona-topic axis. The more general topics (purple, larger marker) have higher rates of exaggeration, and thus caricature, than the specific topics (orange, smaller marker). The uncontroversial (WikiHow, squares) topics have higher rates of caricature than the controversial (ProCon.org, stars) topics. Personas related to political leanings, race/ethnicity, and nonbinary gender broadly have the highest rates of caricature.

A. Wang, J. Morgenstern, J. P. Dickerson, "Large language models cannot replace human participants because they cannot portray identity groups" (2024).

M. Cheng, T. Piccardi, D. Yang, in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023) (Association for Computational Linguistics, 2023).

How might we create models of individuals?

We build models of individuals by providing a detailed description of a persona representing a person.



Personas

Liz Morris is a scientist who researches geology

Julia Vance is an activist who uses a wheelchair

Chris Perez is a gun owner

Mary Rayburn is a grandmother of five from Shreveport who wants to fight for health care.

Bridget Yang is a person who wants to fix student debt

John Hughes is a college student who enjoys anime

Eddie Jackson is a Bernie Sanders supporter

Natalie Wilson is a housewife and grandmother

Phil Johnson is a rational centrist

Sam Thompson is a single father who works two jobs

Paul Smith is a progressive legislator

Linda Brown is an animal rights activist and a vegan

Maggie Washington is a woman living in a small conservative town

Philip Nielsen is a member of the alt-right who is racist against black people

Lucas Pearson is someone who is fed up with the two party system

Travis Howard is a man who works with many marginalized people

Victor Gonzalez is a latino with no political affiliation

Edna Mason is a retired nurse who cares about universal healthcare

Fred Murphy is a social conservative who is forever triggered

Park, J.S., Popowski, L., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2022). Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22). Association for Computing Machinery, New York, NY, USA.

How might we create models of groups?

We build models of groups by providing memories to the models of individuals so that they can interact.

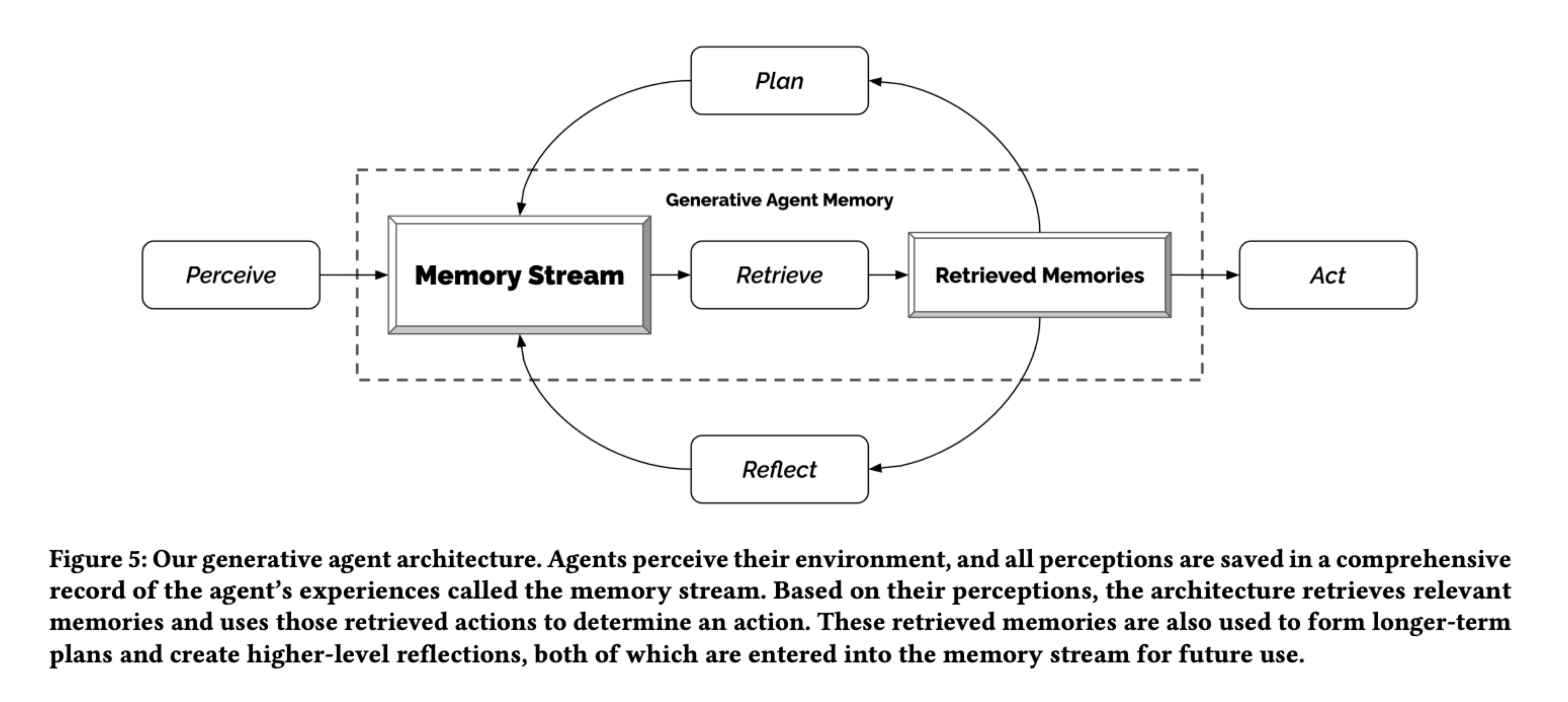


Figure 5: Our generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent's experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories and uses those retrieved actions to determine an action. These retrieved memories are also used to form longer-term plans and create higher-level reflections, both of which are entered into the memory stream for future use.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023).

How might we evaluate the models of individuals and groups?

Idea: convergence and divergence.

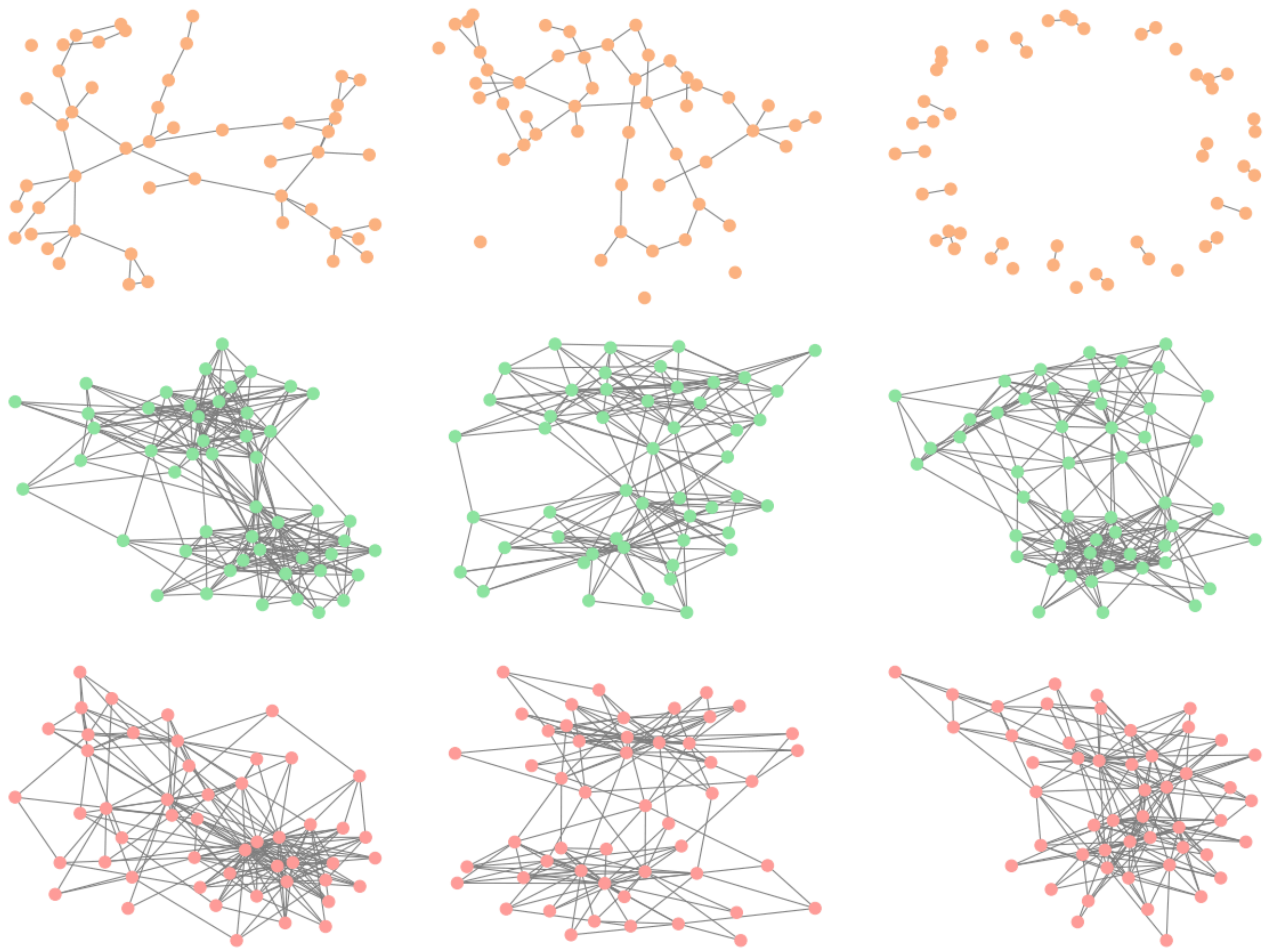


Figure 2: Generated social networks from different prompting methods: Global (top), Local (middle), Sequential (bottom).

Chang, S., Chaszczewicz, A., Wang, E., Josifovska, M., Pierson, E., & Leskovec, J. (2024). LLMs generate structurally realistic social networks but overestimate political homophily. arXiv preprint arXiv:2408.16629.

Bets that we place today.

In summary...

The quantum unit of simulations—individual agents—is an important determinant of a simulation’s success.

Today, many generative AI-based simulations focus on populations.

In summary...

Different level of analysis offer different strengths and weaknesses.

- population: not granular enough

- individuals: too noisy

- groups: might never be predictable

References

- J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023).
- J. von Neumann, Theory of Self-Reproducing Automata, A. W. Burks, Ed. (University of Illinois Press, 1966).
- S. Wolfram, A New Kind of Science (Wolfram Media, 2002).
- SK Card, TP Moran, and A Newell. 1983. The psychology of human-computer interaction. (1983).
- P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). ACM, New York, NY, USA, 175-186.
- Gordon, M.L., Lam, M.S., Park, J.S., Patel, K., Hancock, J.T., Hashimoto, T., & Bernstein, M.S. (2022). Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). Association for Computing Machinery, New York, NY, USA.

References

- Park, J.S., Popowski, L., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2022). Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22). Association for Computing Machinery, New York, NY, USA.
- T. C. Schelling, Dynamic models of segregation. Journal of Mathematical Sociology 1, 143-186 (1971).
- L. P. Argyle et al., Out of one, many: Using language models to simulate human samples. Political Analysis 31, 337-355 (2023).
- A. Ashokkumar, L. Hewitt, I. Ghezae, R. Willer, "Predicting Results of Social Science Experiments Using Large Language Models" (2024).
- J. J. Horton, "Large language models as simulated economic agents: What can we learn from homo silicus?" (2023).
- A. Wang, J. Morgenstern, J. P. Dickerson, "Large language models cannot replace human participants because they cannot portray identity groups" (2024).

References

- M. Cheng, T. Piccardi, D. Yang, in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023) (Association for Computational Linguistics, 2023).
- Chang, S., Chaszczewicz, A., Wang, E., Josifovska, M., Pierson, E., & Leskovec, J. (2024). LLMs generate structurally realistic social networks but overestimate political homophily. arXiv preprint arXiv:2408.16629.

CS 222: AI Agents and Simulations

Stanford University

Joon Sung Park