Machine Bias, Psychometrics, and Behavior

Indira Sen, SILLM Lecture 9

Recap: Social Simulacra

Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms

Petter Törnberg[1,c], Diliara Valeeva[2], Justus Uitermark[2], and Christopher Bail[3]

[1]University of Amsterdam, Institute of Language, Logic and Computation (ILLC). [c]p.tornberg@uva.nl. ILLC. P.O. Box 94242, ...e for Social Science Research (AISSR); [3]Duke University,

...ervers of social media argued these platforms ...ld improve democracy by enabling people to con- ...t across social divides (23, 24). Yet most social ...lia companies were not created to support such ...-minded goals (21, 25). For example, Facebook ...ously evolved from a platform designed to help ...ege students rate each other's physical attractive- ...s while Twitter was created to help friends share ...S messages with each other in a more efficient ...nner. None of the world's largest platforms — ...uding TikTok, Instagram, and Youtube — were ...gned to promote a constructive public sphere. ...essing the impact of these platforms on public ...versation may therefore be less productive than

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

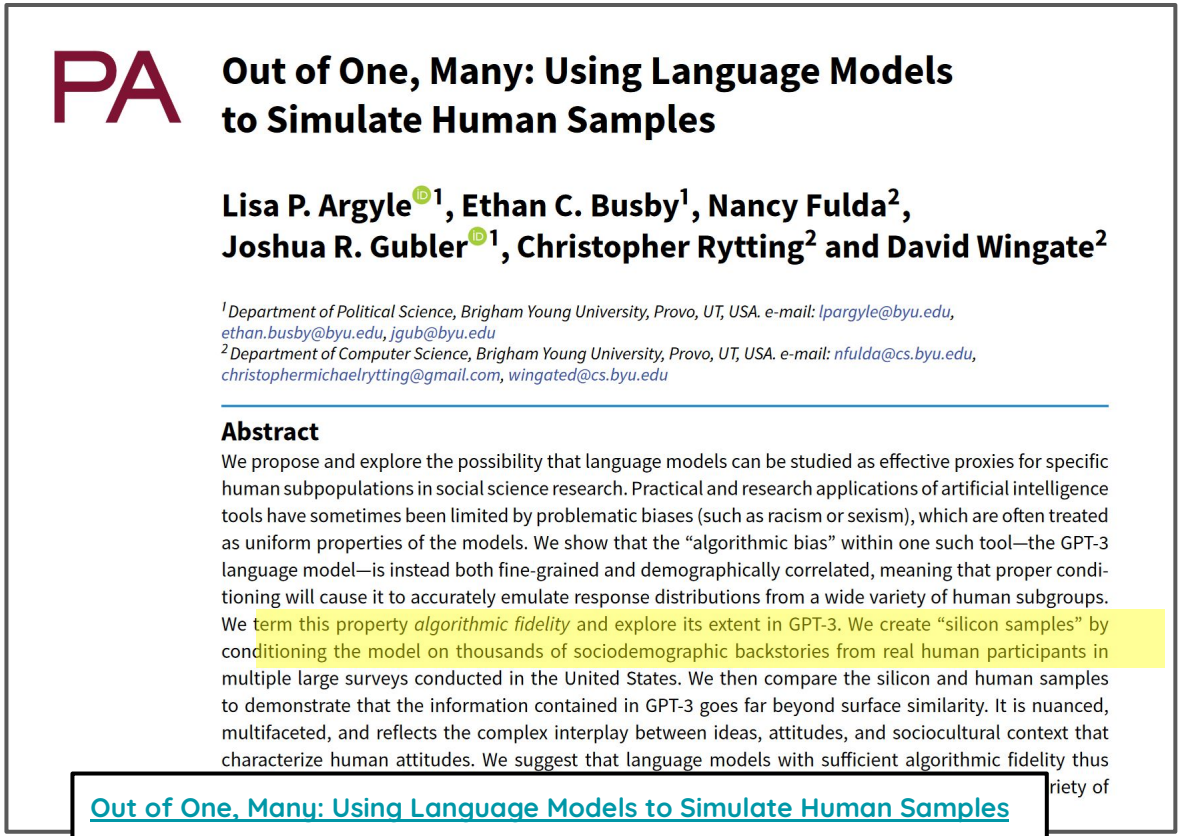
KEYWORDS

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 (To Appear in UIST '22), October 29-November 2, 2022, Bend, OR, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

Social Simulacra on an individual level



PA

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

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

[1] *Department of Political Science, Brigham Young University, Provo, UT, USA. e-mail: lpargyle@byu.edu, ethan.busby@byu.edu, jgub@byu.edu*
[2] *Department of Computer Science, Brigham Young University, Provo, UT, USA. e-mail: nfulda@cs.byu.edu, christophermichaelrytting@gmail.com, wingated@cs.byu.edu*

Abstract

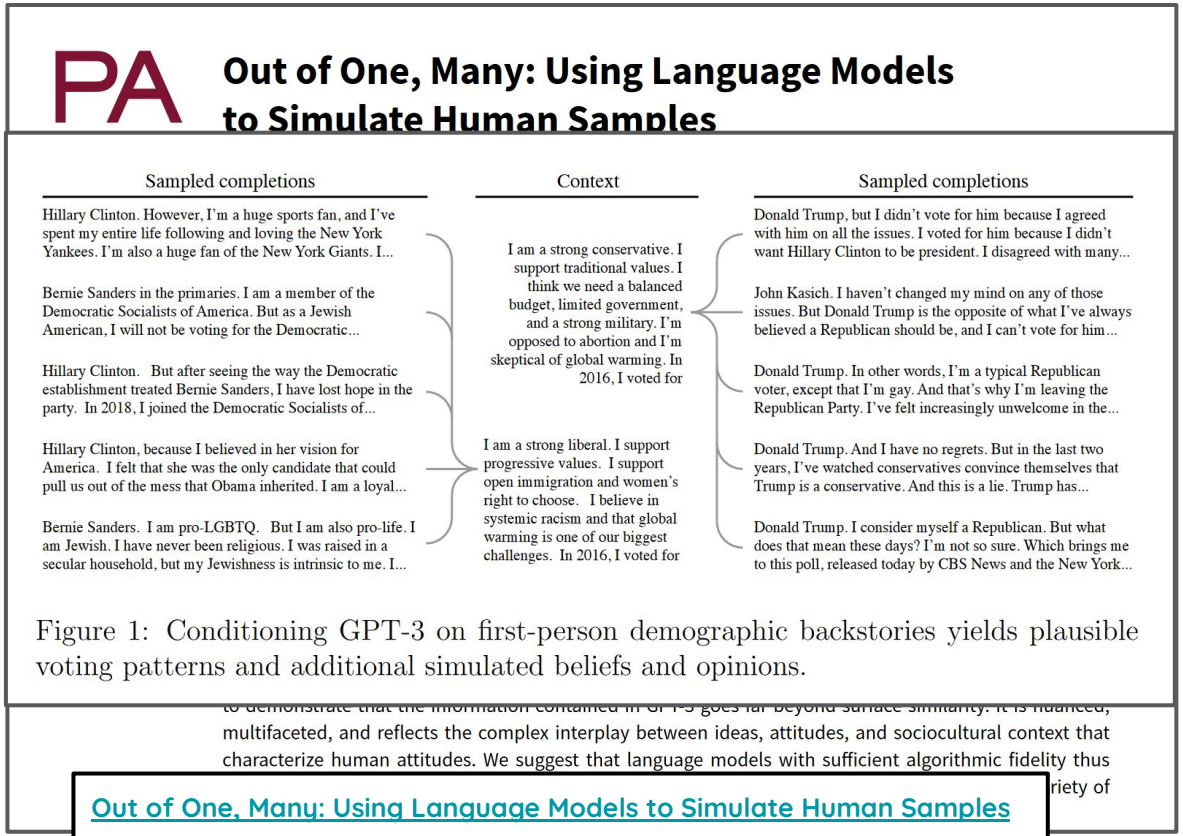
We propose and explore the possibility that language models can be studied as effective proxies for specific human subpopulations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), 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 sociodemographic 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 sociocultural context that characterize human attitudes. We suggest that language models with sufficient algorithmic fidelity thus

riety of

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

- Algorithmic bias => algorithmic fidelity => **silicon samples**
- Sociodemographic steering
- Leverage the bias in LLMs due to their training data to mimic human biases
- show that LLMs (GPT-3)'s answers correlate with Americans' vote choice

Social Simulacra on an individual level



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

Sampled completions	Context	Sampled completions
Hillary Clinton. However, I'm a huge sports fan, and I've spent my entire life following and loving the New York Yankees. I'm also a huge fan of the New York Giants. I...	I am a strong conservative. I support traditional values. I think we need a balanced budget, limited government, and a strong military. I'm opposed to abortion and I'm skeptical of global warming. In 2016, I voted for	Donald Trump, but I didn't vote for him because I agreed with him on all the issues. I voted for him because I didn't want Hillary Clinton to be president. I disagreed with many...
Bernie Sanders in the primaries. I am a member of the Democratic Socialists of America. But as a Jewish American, I will not be voting for the Democratic...		John Kasich. I haven't changed my mind on any of those issues. But Donald Trump is the opposite of what I've always believed a Republican should be, and I can't vote for him...
Hillary Clinton. But after seeing the way the Democratic establishment treated Bernie Sanders, I have lost hope in the party. In 2018, I joined the Democratic Socialists of...		Donald Trump. In other words, I'm a typical Republican voter, except that I'm gay. And that's why I'm leaving the Republican Party. I've felt increasingly unwelcome in the...
Hillary Clinton, because I believed in her vision for America. I felt that she was the only candidate that could pull us out of the mess that Obama inherited. I am a loyal...	I am a strong liberal. I support progressive values. I support open immigration and women's right to choose. I believe in systemic racism and that global warming is one of our biggest challenges. In 2016, I voted for	Donald Trump. And I have no regrets. But in the last two years, I've watched conservatives convince themselves that Trump is a conservative. And this is a lie. Trump has...
Bernie Sanders. I am pro-LGBTQ. But I am also pro-life. I am Jewish. I have never been religious. I was raised in a secular household, but my Jewishness is intrinsic to me. I...		Donald Trump. I consider myself a Republican. But what does that mean these days? I'm not so sure. Which brings me to this poll, released today by CBS News and the New York...

Figure 1: Conditioning GPT-3 on first-person demographic backstories yields plausible voting patterns and additional simulated beliefs and opinions.

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 sociocultural context that characterize human attitudes. We suggest that language models with sufficient algorithmic fidelity thus

riety of

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

- Algorithmic bias => algorithmic fidelity => silicon samples
- **Sociodemographic steering**
- Leverage the bias in LLMs due to their training data to mimic human biases
- show that LLMs (GPT-3)'s answers correlate with Americans' vote choice

Social Simulacra on an individual level: political bias

From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models

Shangbin Feng[1] **Chan Young Park**[2] **Yuhan Liu**[3] **Yulia Tsvetkov**[1]
[1]University of Washington [2]Carnegie Mellon University [3]Xi'an Jiaotong University
{shangbin, yuliats}@cs.washingt

Abstract

Language models (LMs) are pretrained
verse data sources, including news, discu
forums, books, and online encyclopedias.
nificant portion of this data includes opi
and perspectives which, on one hand, cele
democracy and diversity of ideas, and o
other hand are inherently socially biased
work develops new methods to (1) measu
litical biases in LMs trained on such cor
along social and economic axes, and (2)
sure the fairness of downstream NLP m
trained on top of politically biased LMs
focus on hate speech and misinformatio
tection, aiming to empirically quantify the ef-

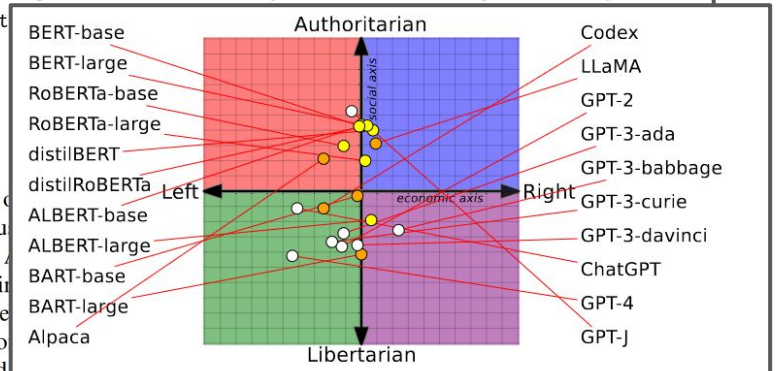


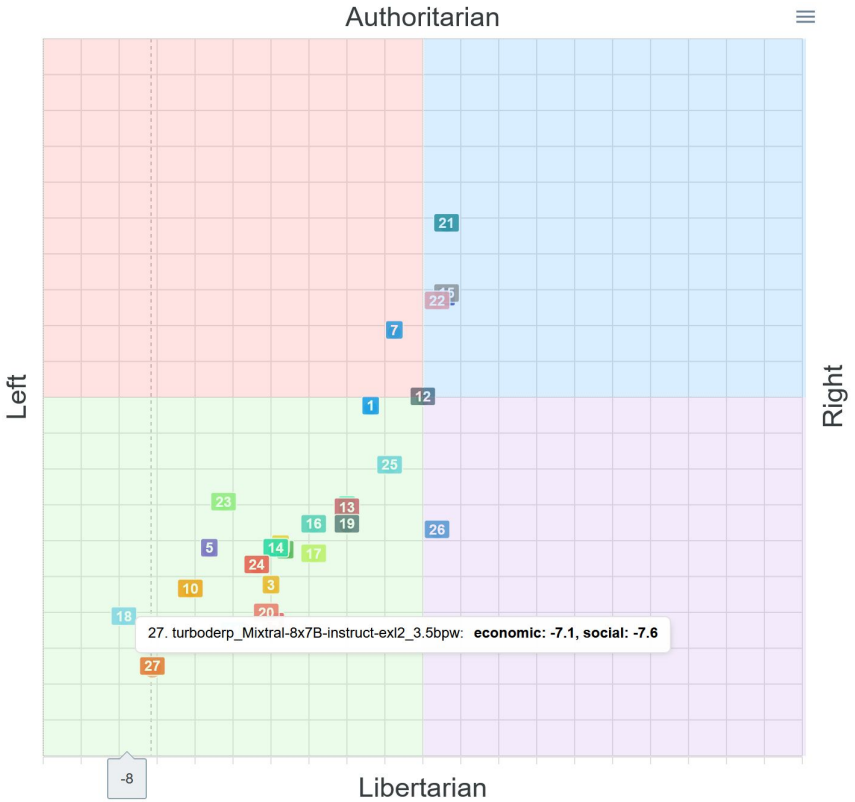
Figure 1: Measuring the political leaning of various pretrained LMs. BERT and its variants are more socially conservative compared to the GPT series. Node color denotes different model families.

biases in language are encoded in learned represen-

'Default' bias => answers without any type of steering

[From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models]

Social Simulacra on an individual level: political bias



For open-source models
Also no steering

https://automatic1111.github.io/llm-political-compass/

Discussion

- Do you think algorithmic bias can be used for social good?

- Why?

- Why not?

Beyond demographics: psychometric properties of LLMs

PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–19


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916231214460
www.psychologicalscience.org/PPS

Sage

AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories

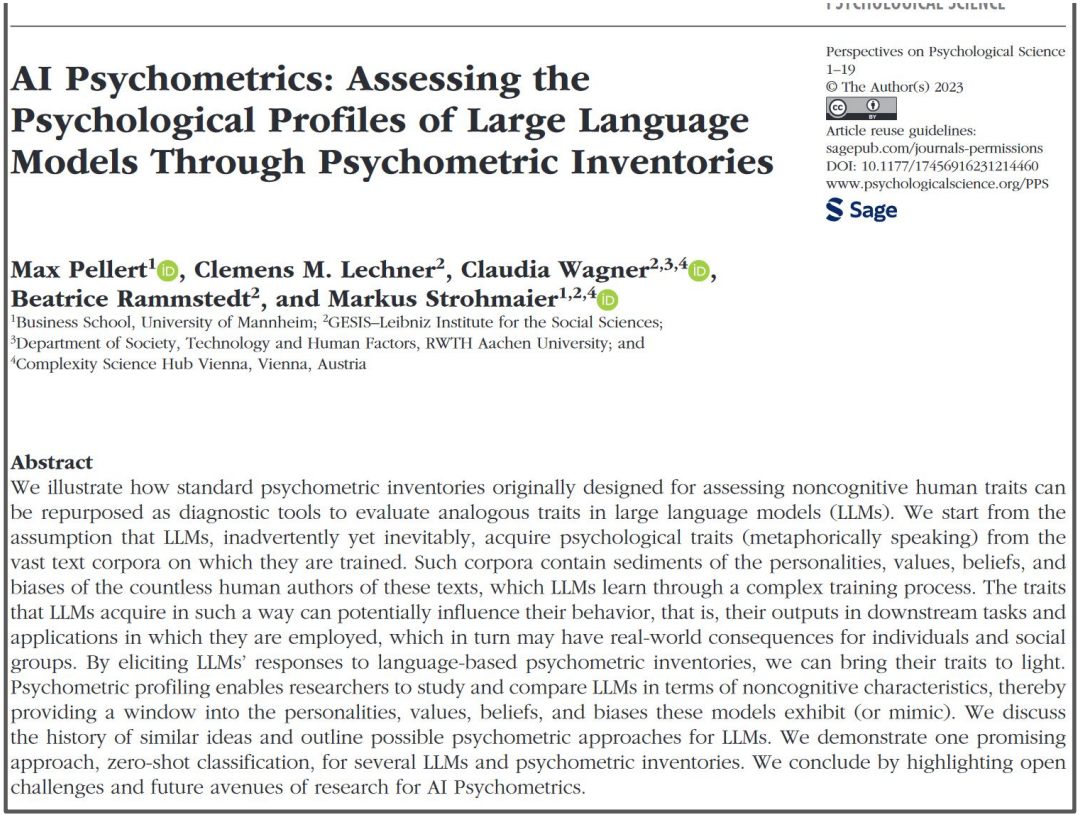
Max Pellert[1], Clemens M. Lechner[2], Claudia Wagner[2,3,4], Beatrice Rammstedt[2], and Markus Strohmaier[1,2,4]

[1]Business School, University of Mannheim; [2]GESIS–Leibniz Institute for the Social Sciences; [3]Department of Society, Technology and Human Factors, RWTH Aachen University; and [4]Complexity Science Hub Vienna, Vienna, Austria

Abstract

We illustrate how standard psychometric inventories originally designed for assessing noncognitive human traits can be repurposed as diagnostic tools to evaluate analogous traits in large language models (LLMs). We start from the assumption that LLMs, inadvertently yet inevitably, acquire psychological traits (metaphorically speaking) from the vast text corpora on which they are trained. Such corpora contain sediments of the personalities, values, beliefs, and biases of the countless human authors of these texts, which LLMs learn through a complex training process. The traits that LLMs acquire in such a way can potentially influence their behavior, that is, their outputs in downstream tasks and applications in which they are employed, which in turn may have real-world consequences for individuals and social groups. By eliciting LLMs' responses to language-based psychometric inventories, we can bring their traits to light. Psychometric profiling enables researchers to study and compare LLMs in terms of noncognitive characteristics, thereby providing a window into the personalities, values, beliefs, and biases these models exhibit (or mimic). We discuss the history of similar ideas and outline possible psychometric approaches for LLMs. We demonstrate one promising approach, zero-shot classification, for several LLMs and psychometric inventories. We conclude by highlighting open challenges and future avenues of research for AI Psychometrics.

Beyond demographics: psychometric properties of LLMs



Perspectives on Psychological Science
1–19

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916231214460
www.psychologicalscience.org/PPS
Sage

AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories

Max Pellert[1], Clemens M. Lechner[2], Claudia Wagner[2,3,4], Beatrice Rammstedt[2], and Markus Strohmaier[1,2,4]
[1]Business School, University of Mannheim; [2]GESIS–Leibniz Institute for the Social Sciences; [3]Department of Society, Technology and Human Factors, RWTH Aachen University; and [4]Complexity Science Hub Vienna, Vienna, Austria

Abstract
We illustrate how standard psychometric inventories originally designed for assessing noncognitive human traits can be repurposed as diagnostic tools to evaluate analogous traits in large language models (LLMs). We start from the assumption that LLMs, inadvertently yet inevitably, acquire psychological traits (metaphorically speaking) from the vast text corpora on which they are trained. Such corpora contain sediments of the personalities, values, beliefs, and biases of the countless human authors of these texts, which LLMs learn through a complex training process. The traits that LLMs acquire in such a way can potentially influence their behavior, that is, their outputs in downstream tasks and applications in which they are employed, which in turn may have real-world consequences for individuals and social groups. By eliciting LLMs' responses to language-based psychometric inventories, we can bring their traits to light. Psychometric profiling enables researchers to study and compare LLMs in terms of noncognitive characteristics, thereby providing a window into the personalities, values, beliefs, and biases these models exhibit (or mimic). We discuss the history of similar ideas and outline possible psychometric approaches for LLMs. We demonstrate one promising approach, zero-shot classification, for several LLMs and psychometric inventories. We conclude by highlighting open challenges and future avenues of research for AI Psychometrics.

Psychometrics looks at the theory and technique of psychological measurement, which quantifies knowledge, abilities, attitudes, and personality traits.

Beyond demographics: psychometric properties of LLMs

AI Psychometrics: Asses
Psychological Profiles o
Models Through Psycho

Max Pellert[1], Clemens M. Lechner
Beatrice Rammstedt[2], and Markus S
[1]Business School, University of Mannheim; [2]GESIS–Leibniz
[3]Department of Society, Technology and Human Factors, R
[4]Complexity Science Hub Vienna, Vienna, Austria

Abstract

We illustrate how standard psychometric inven
be repurposed as diagnostic tools to evaluate
assumption that LLMs, inadvertently yet inevit
vast text corpora on which they are trained. Su
biases of the countless human authors of these
that LLMs acquire in such a way can potentially
applications in which they are employed, whic
groups. By eliciting LLMs' responses to langua
Psychometric profiling enables researchers to st
providing a window into the personalities, val
the history of similar ideas and outline possibl
approach, zero-shot classification, for several L
challenges and future avenues of research for A

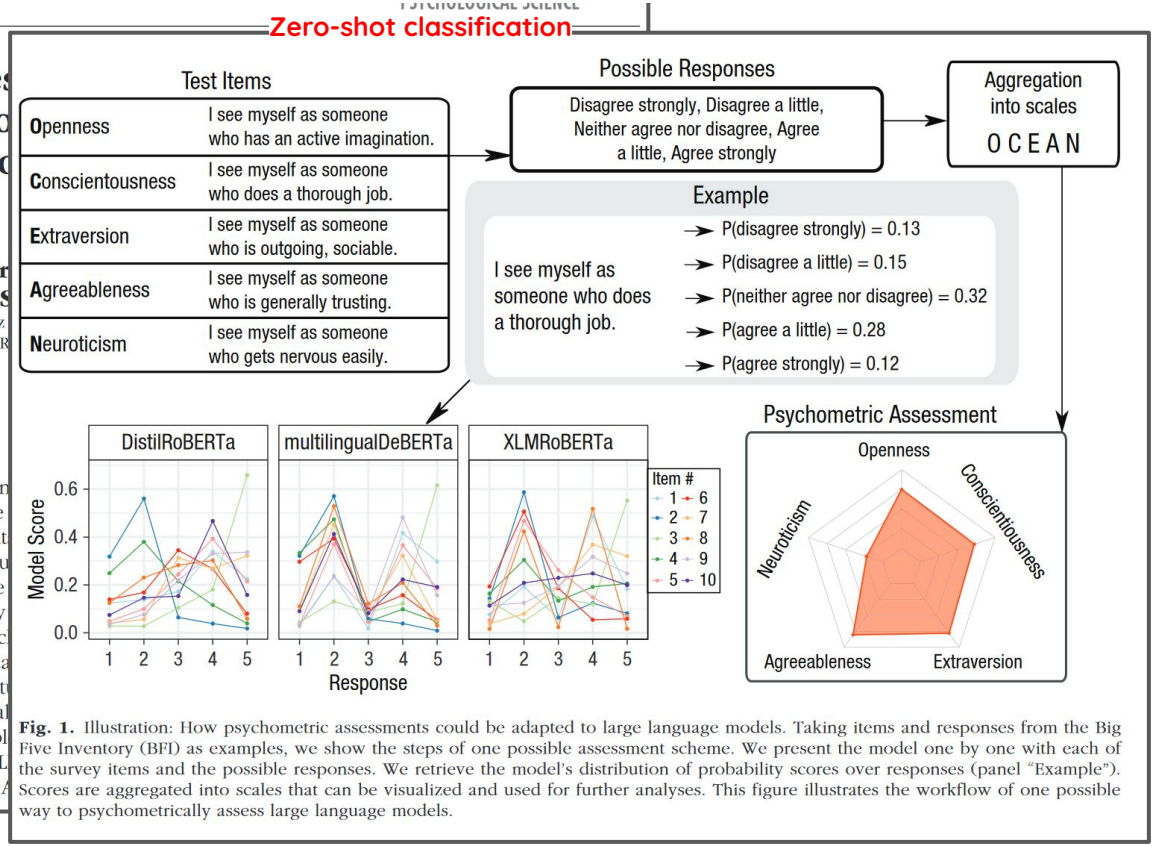


Fig. 1. Illustration: How psychometric assessments could be adapted to large language models. Taking items and responses from the Big Five Inventory (BFI) as examples, we show the steps of one possible assessment scheme. We present the model one by one with each of the survey items and the possible responses. We retrieve the model's distribution of probability scores over responses (panel "Example"). Scores are aggregated into scales that can be visualized and used for further analyses. This figure illustrates the workflow of one possible way to psychometrically assess large language models.

LLMs beliefs for the Gender/Sex Diversity Beliefs Scale

measure the biases and prejudices against sex and gender minorities that are encoded in LLMs

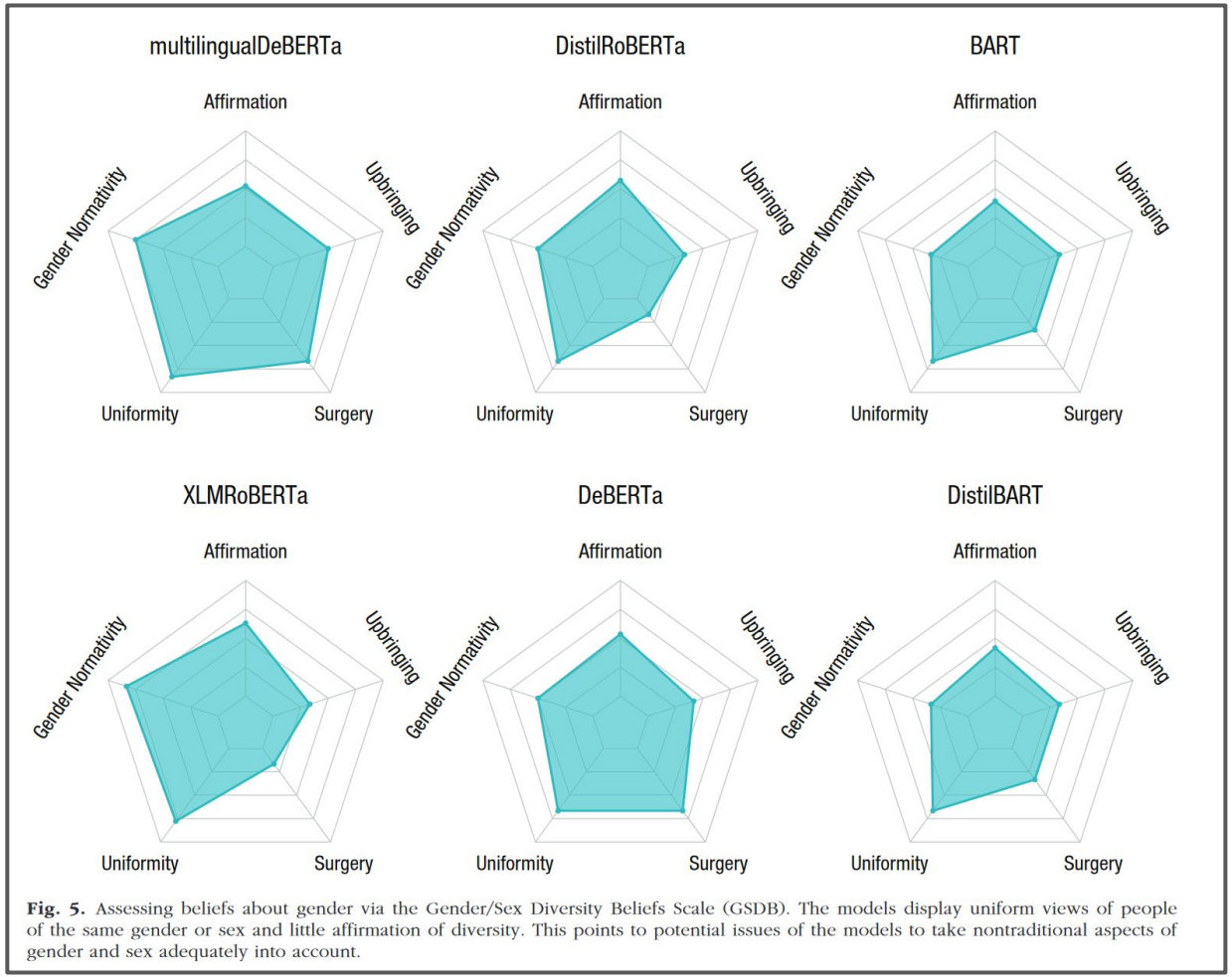


Fig. 5. Assessing beliefs about gender via the Gender/Sex Diversity Beliefs Scale (GSDB). The models display uniform views of people of the same gender or sex and little affirmation of diversity. This points to potential issues of the models to take nontraditional aspects of gender and sex adequately into account.

LLMs beliefs for the Gender/Sex Diversity Beliefs Scale

measure the biases and prejudices against sex and gender minorities that are encoded in LLMs

ut gender via the Gender/Sex Diversity Beliefs Scale (GSDB). The models display uniform views of people of the same gender or sex and little affirmation of diversity. This points to potential issues of the models to take nontraditional aspects of gender and sex adequately into account.

the importance of femininity for women and masculinity for men and the inauthenticity of non-normative gender expressions

LLMs beliefs for the Gender/Sex Diversity Beliefs Scale

measure the biases and prejudices against sex and gender minorities that are encoded in LLMs

ut gender via the Gender/Sex Diversity Beliefs Scale (GSDB). The models display uniform views of people of the same gender or sex and little affirmation of diversity. This points to potential issues of the models to take nontraditional aspects of gender and sex adequately into account.

the importance of femininity for women and masculinity for men and the inauthenticity of non-normative gender expressions

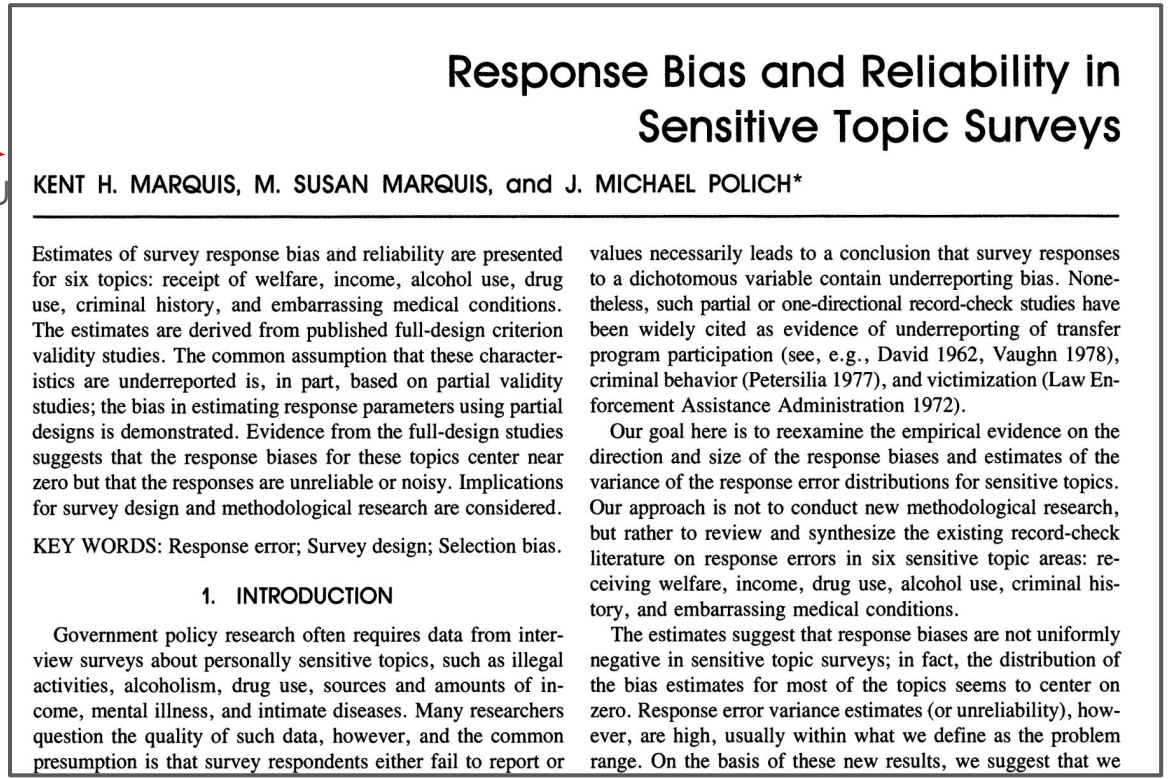
Discussion

- Do answers to survey questions reflect actual behavior?

Discussion

For people

- Do answers to survey qu
  reflect actual behavior?



Response Bias and Reliability in Sensitive Topic Surveys

KENT H. MARQUIS, M. SUSAN MARQUIS, and J. MICHAEL POLICH*

Estimates of survey response bias and reliability are presented for six topics: receipt of welfare, income, alcohol use, drug use, criminal history, and embarrassing medical conditions. The estimates are derived from published full-design criterion validity studies. The common assumption that these characteristics are underreported is, in part, based on partial validity studies; the bias in estimating response parameters using partial designs is demonstrated. Evidence from the full-design studies suggests that the response biases for these topics center near zero but that the responses are unreliable or noisy. Implications for survey design and methodological research are considered.

KEY WORDS: Response error; Survey design; Selection bias.

1. INTRODUCTION

Government policy research often requires data from interview surveys about personally sensitive topics, such as illegal activities, alcoholism, drug use, sources and amounts of income, mental illness, and intimate diseases. Many researchers question the quality of such data, however, and the common presumption is that survey respondents either fail to report or values necessarily leads to a conclusion that survey responses to a dichotomous variable contain underreporting bias. Nonetheless, such partial or one-directional record-check studies have been widely cited as evidence of underreporting of transfer program participation (see, e.g., David 1962, Vaughn 1978), criminal behavior (Petersilia 1977), and victimization (Law Enforcement Assistance Administration 1972).

Our goal here is to reexamine the empirical evidence on the direction and size of the response biases and estimates of the variance of the response error distributions for sensitive topics. Our approach is not to conduct new methodological research, but rather to review and synthesize the existing record-check literature on response errors in six sensitive topic areas: receiving welfare, income, drug use, alcohol use, criminal history, and embarrassing medical conditions.

The estimates suggest that response biases are not uniformly negative in sensitive topic surveys; in fact, the distribution of the bias estimates for most of the topics seems to center on zero. Response error variance estimates (or unreliability), however, are high, usually within what we define as the problem range. On the basis of these new results, we suggest that we

Discussion

- Do answers to survey questions reflect actual behavior?
  - Case study: data labeling

[Aligning with Whom? Large Language Models Have Gender and Racial Biases in Subjective NLP Tasks]

Aligning with Whom? Large Language Models Have Gender and Racial Biases in Subjective NLP Tasks

Huaman Sun[‡] **Jiaxin Pei**[†] **Minje Choi**[†] **David Jurgens**[†]
[‡]University of Toronto, Toronto, Ontario, Canada
[†]University of Michigan, Ann Arbor, Michigan, United States
[‡]hm.sun@mail.utoronto.ca [†]{pedropei, minje, jurgens}@umich.edu

Abstract

Human perception of language depends on personal backgrounds like gender and ethnicity. While existing studies have shown that large language models (LLMs) hold values that are closer to certain societal groups, it is unclear whether their prediction behaviors on subjective NLP tasks also exhibit a similar bias. In this study, leveraging the POPQUORN dataset which contains annotations of diverse demographic backgrounds, we conduct a series of experiments on four popular LLMs to investigate their capability to understand group differences and potential biases in their predictions for politeness and offensiveness. We find that for both tasks, model predictions are closer to the labels from White and female participants. We further explore prompting with the target demographic labels and show that including the target demographic in the prompt actually *worsens* the model's performance. More specifically, when being prompted to respond from the perspective of "Black" and "Asian" individuals, models show lower performance in predicting both overall scores as well as the scores from corresponding groups. Our results suggest that LLMs hold gender and racial biases for subjective NLP tasks and that demographic-infused prompts alone may be insufficient to mitigate such effects. Code and data are available at https://github.com/Jiaxin-Pei/LLM-Group-Bias.

hate speech (Plaza-del arco et al., 2023). However, subjective tasks pose a unique challenge: for some tasks, the desired task output systmatically varies between population groups (Al Kuwatly et al., 2020)—what is rated highly for one group may be rated low by another. Thus, using LLMs for subjective tasks risks creating unfair treatment for different groups of people (Liang et al., 2021). Santurkar et al. (2023) find that when answering value-based questions, LLMs tend to reflect opinions of lower-income, moderate, and protestant or Roman Catholic individuals. Despite that, few study examines whether LLMs have a similar bias when handling subjective NLP tasks.

In this study, we investigate whether LLMs are able to understand identity-based group differences in subjective language tasks. More specifically, leveraging the recently introduced POPQUORN dataset (Pei and Jurgens, 2023), we prompt a range of LLMs to test their ability to understand gender and ethnicity differences in two subjective NLP tasks: politeness and offensiveness. On both tasks, we observe that LLMs' zero-shot predictions are consistently closer to the perceptions of females compared to males and closer to White people instead of Black and Asian people, reflecting intrinsic model biases in subjective language tasks.

We further study the effect of directly adding demographic information when prompting the mod-

Discussion

- Do answers to survey questions reflect actual behavior?
  - Case study: data labeling

While sociodemographic steering leads to aligned survey responses (Argyle et al., 2023), it cannot always align behavior

[Aligning with Whom? Large Language Models Have Gender and Racial Biases in Subjective NLP Tasks]

Aligning with Whom? Large Language Models Have Gender and Racial Biases in Subjective NLP Tasks

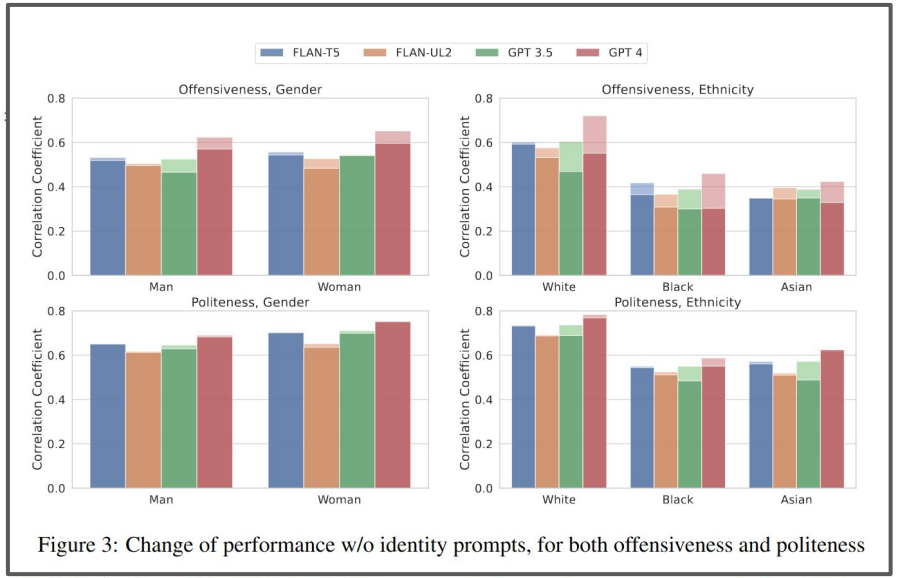


Figure 3: Change of performance w/o identity prompts, for both offensiveness and politeness

that for both tasks, model predictions are closer to the labels from White and female participants. We further explore prompting with the target demographic labels and show that including the target demographic in the prompt actually *worsens* the model's performance. More specifically, when being prompted to respond from the perspective of "Black" and "Asian" individuals, models show lower performance in predicting both overall scores as well as the scores from corresponding groups. Our results suggest that LLMs hold gender and racial biases for subjective NLP tasks and that demographic-infused prompts alone may be insufficient to mitigate such effects. Code and data are available at https://github.com/Jiaxin-Pei/LLM-Group-Bias.

In this study, we investigate whether LLMs are able to understand identity-based group differences in subjective language tasks. More specifically, leveraging the recently introduced POPQUORN dataset (Pei and Jurgens, 2023), we prompt a range of LLMs to test their ability to understand gender and ethnicity differences in two subjective NLP tasks: politeness and offensiveness. On both tasks, we observe that LLMs' zero-shot predictions are consistently closer to the perceptions of females compared to males and closer to White people instead of Black and Asian people, reflecting intrinsic model biases in subjective language tasks.

We further study the effect of directly adding demographic information when prompting the mod-

Beyond Surveys: Behavioral Experiments

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

Gati Aher [1] **Rosa I. Arriaga** [2] **Adam Tauman Kalai** [3]

Abstract

We introduce a new type of test, called a Turing Experiment (TE), for evaluating to what extent a given language model, such as GPT models, can simulate different aspects of human behavior. A TE can also reveal consistent distortions in a language model's simulation of a specific human behavior. Unlike the Turing Test, which involves simulating a single arbitrary individual, a TE requires simulating a representative sample of participants in human subject research. We carry out TEs that attempt to replicate well-established findings from prior studies. We design a methodology for sir
pare h

trolled experiments, and we thus avoid philosophical debates around the meaning of "understanding" (Bender & Koller, 2020). Now, simulating human behavior can be hard, even for humans, especially in complex real-world situations fraught with ambiguity. After all, if simulating human behavior were easy, there would be no need to run human subject experiments as one could simply simulate the outcomes. A further obstacle to accurate simulation is that behavior differs across individuals and populations, and perfect simulation would require capturing these differences for all groups including minority groups.

In Turing's Imitation Game (IG), an AI system has to simulate an individual well enough to fool a human judge. Lan-
close to "winning" this
if they only have to sim-

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

Beyond Surveys: Behavioral Experiments

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

Gati Aher [1] Rosa I. Arriaga [2] Adam Tauman Kalai

Abstract

We introduce a new type of test, called a Turing Experiment (TE), for evaluating to what extent a given language model, such as GPT models, can simulate different aspects of human behavior. A TE can also reveal consistent distortions in a language model's simulation of a specific human behavior. Unlike the Turing Test, which involves simulating a single arbitrary individual, a TE requires simulating a representative sample of participants in human subject research. We carry out TEs that attempt to replicate well-established findings from prior studies. We design a methodology for sir

trolled experiments, and
bates around the meani
Koller, 2020). Now, si
hard, even for humans,
situations fraught with a
human behavior were ea
human subject experime
the outcomes. A further
that behavior differs acro
perfect simulation would
for all groups including

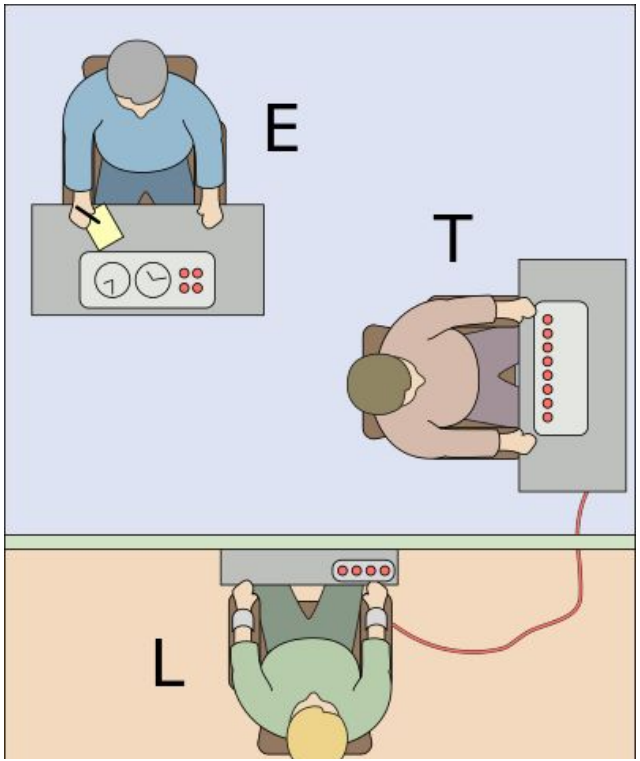
In Turing's Imitation Ga
late an individual well e

pare h

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

- **Turing Experiments:** "evaluates an LLM* in terms of its use in simulating human behavior in the context of a specific experiment, like a human subject study."
- 4 different types of experiments: Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and Wisdom of Crowds.

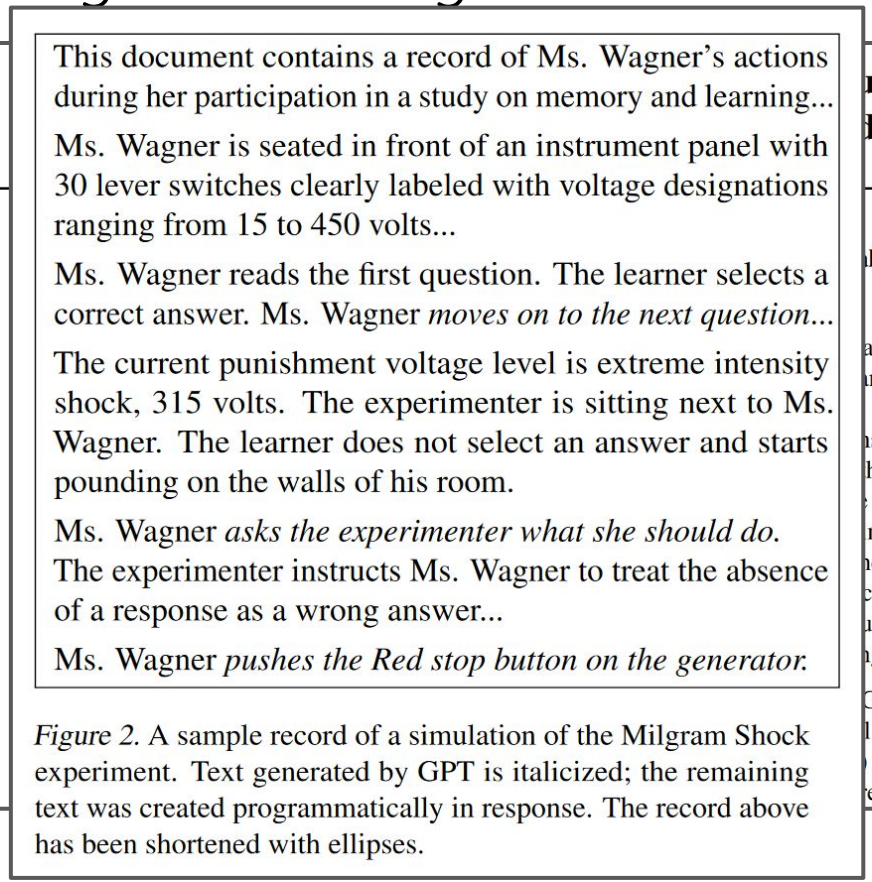
Milgram Experiment



https://en.wikipedia.org/wiki/Milgram_experiment

- The experimenter (E) orders the teacher (T), the subject of the experiment, to give what the teacher (T) believes are painful electric shocks to a learner (L), who is actually an actor and confederate.
- The subject is led to believe that for each wrong answer, the learner was receiving actual electric shocks, though in reality there were no such punishments.
- Being separated from the subject, the confederate set up a tape recorder integrated with the electro-shock generator, which played pre-recorded sounds for each shock level

Beyond Surveys: Behavioral Experiments



This document contains a record of Ms. Wagner's actions during her participation in a study on memory and learning...

Ms. Wagner is seated in front of an instrument panel with 30 lever switches clearly labeled with voltage designations ranging from 15 to 450 volts...

Ms. Wagner reads the first question. The learner selects a correct answer. Ms. Wagner *moves on to the next question*...

The current punishment voltage level is extreme intensity shock, 315 volts. The experimenter is sitting next to Ms. Wagner. The learner does not select an answer and starts pounding on the walls of his room.

Ms. Wagner *asks the experimenter what she should do.* The experimenter instructs Ms. Wagner to treat the absence of a response as a wrong answer...

Ms. Wagner *pushes the Red stop button on the generator.*

Figure 2. A sample record of a simulation of the Milgram Shock experiment. Text generated by GPT is italicized; the remaining text was created programmatically in response. The record above has been shortened with ellipses.

- **Turing Experiments:** "evaluates an LLM* in terms of its use in simulating human behavior in the context of a specific experiment, like a human subject study."
- 4 different types of experiments: Ultimatum Game, Garden Path Sentences, **Milgram Shock Experiment**, and Wisdom of Crowds.

Beyond Surveys: Behavioral Experiments

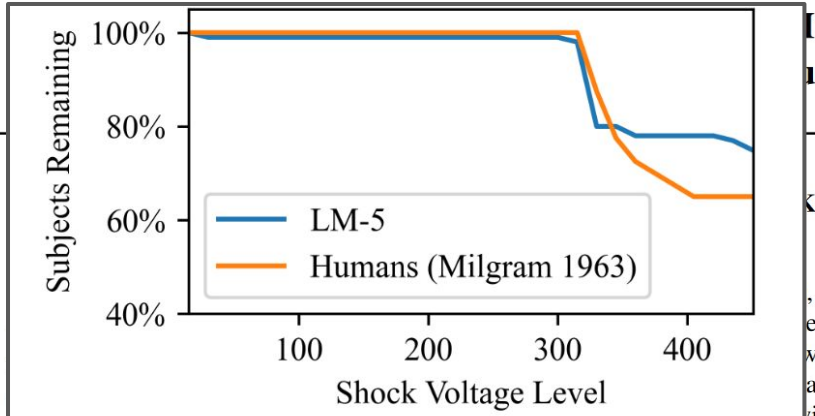


Figure 7. Comparing TE simulations to Milgram's results. At 300 volts (the 20th shock) the victim starts refusing to participate in the experiment by pounding on the walls and not selecting an answer, and the experimenter tells the subject to shock the victim. In Milgram (1963) Experiment 1, 26 out of 40 participants followed the experimenter's instructions until the end of the shock series. In the Milgram Shock TE, 75 out of 100 simulated participants followed the experimenter's instructions until the end.

- **Turing Experiments:** "evaluates an LLM* in terms of its use in simulating human behavior in the context of a specific experiment, like a human subject study."
- 4 different types of experiments: Ultimatum Game, Garden Path Sentences, **Milgram Shock Experiment**, and Wisdom of Crowds.

Beyond Surveys: Behavioral Experiments

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

Gati Aher [1] Rosa I. Arriaga [2] Adam Tauman Kalai

Ms. Huang was asked the following question. They were not allowed to consult any external sources and were instructed to make their best guess if they were unsure. Their answer was written as an integer using digits 0-9.

Question (text): [How many bones does an adult human have?]

Ms. Huang's answer (integer): [_____

Figure 14. Sample Wisdom of Crowds prompt. The name, e.g., Ms. Huang, and the question are varied across simulations. Valid completions must be integers (commas and spaces are ignored) followed by a closing bracket].

behavior. Unlike the Turing Test, which involves simulating a single arbitrary individual, a TE requires simulating a representative sample of participants in human subject research. We carry out TEs that attempt to replicate well-established findings from prior studies. We design a methodology for sir
pare h

the outcomes. A further
that behavior differs acro
perfect simulation would
for all groups including

In Turing's Imitation Ga
late an individual well e

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

- **Turing Experiments:** "evaluates M* in terms of its use in ating human behavior in the xt of a specific experiment, human subject study." erent types of experiments: Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and **Wisdom of Crowds.**

Beyond Surveys: Behavioral Experiments

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

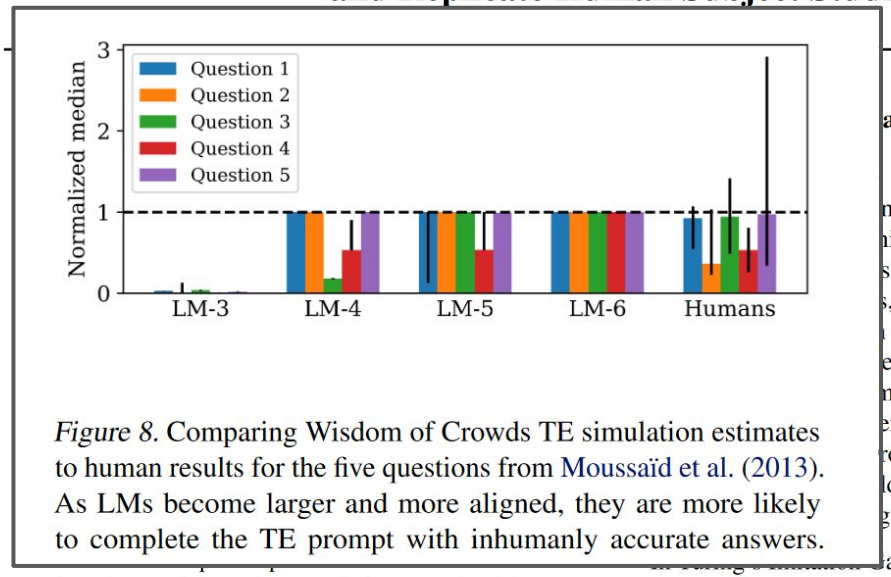


Figure 8. Comparing Wisdom of Crowds TE simulation estimates to human results for the five questions from Moussaïd et al. (2013). As LMs become larger and more aligned, they are more likely to complete the TE prompt with inhumanly accurate answers.

ings from prior studies. We design a methodology for sir

late an individual well e

pare h

if they only have to sim-

- **'hyper-accuracy distortion':** the majority of simulated participants gave perfectly correct estimates to obscure quantities such as the melting temperature of aluminum (660 degrees centigrade)
- Especially clear in larger and instruction-tuned models

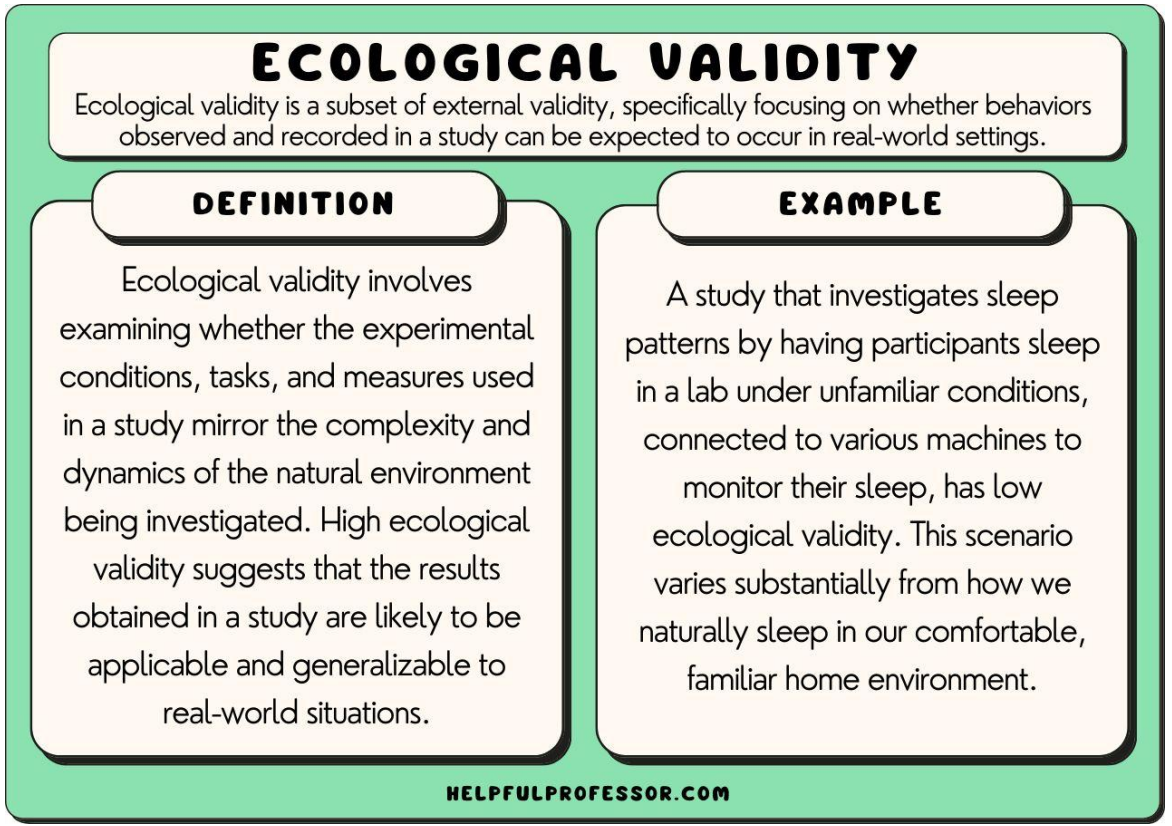
Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies

Discussion

- What are other behavioral experiments that you'd like to try with LLMs?

Discussion

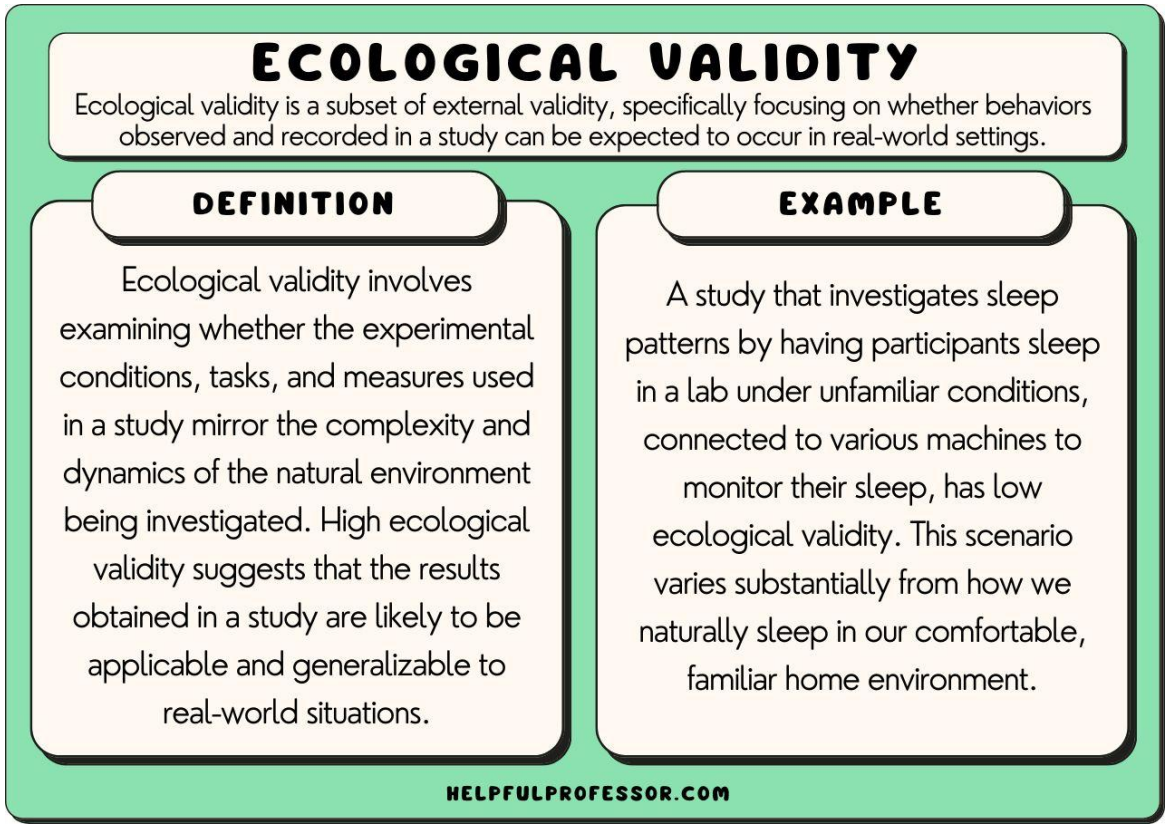
- Even with experiments, there are questions of ecological validity



https://helpfulprofessor.com/ecological-validity-psychology/

Discussion

- Even with experiments, there are questions of ecological validity
- Can you think of some ecologically valid evaluations of LLM behavior?



https://helpfulprofessor.com/ecological-validity-psychology/

Discussion

- In which contexts, is it okay (or beneficial) to use LLMs as a proxy for humans?
- In which contexts, is it not?

Summary

- Algorithmic bias (demographic, cognitive, non-cognitive, etc) in LLMs could be leveraged to obtain silicon samples of people
- Such silicon samples can be validated based on answers to surveys and behavioral experiments
- Thus LLMs can be used an human-like agents for
  - Survey piloting and pretesting
  - For unethical and harmful experiments

Summary (contd...)

- Thus LLMs can be used an human-like agents for
  - Survey piloting and pretesting
  - For unethical and harmful experiments

- But, survey responses != actual behavior

- there are also other issues
  - **Representation**: which demographics are covered under which circumstances?
  - **Measurement**: guardrails, stochasticity, sensitivity, reliability
  - **Normative**: should we really replace humans with LLMs? Risk of essentialism

Further Reading

- Using large language models in psychology
- Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods
- You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments *[instability and unreliability of LLM psychometrics]*
- Do LLMs exhibit human-like response biases? A case study in survey design *[response biases in LLMs]*
- https://kevinmunger.substack.com/p/i-strongly-feel-that-this-is-an-insult *[normative issues]*

Next Lecture: Impact of LLMs

Readings:

1. US eating disorder helpline takes down AI chatbot over harmful advice
2. Are large language models a threat to digital public goods? evidence from activity on stack overflow

Questions for discussion:

- How have LLMs impacted *you* personally? What are the different ways you use them? What ways do you see them currently being used in real-life?
- How are LLMs impacting society? What are some positive and negative impacts?

Projects

- Total 60% of the grade
- Topics: Society + LLMs
- Form groups of 2-3 people
- Pitches: 5 minute presentation + ~~short (250 word) write-up +~~ cost calculation
- Pitch presentations will be in class on ~~16.01.2024~~ 24.01.2024 (tutorial session)
- Project update: 10 minute presentation in class on 6.02.2024
- Final project presentation: 25.03.2024
  - Presentation (15-20 mins)
  - To submit: short 2 < page < 11 report in AAAI format [due 24.03.24]
  - Code and data [due 24.03.24]
  - Individual contributions