

At the Intersection of **NLP** and **Survey Methodology** : Potentials, Challenges, and Provocations

Indira Sen

CS3 Meeting for Computational Survey and Social Science
22.05.25

NLP and Survey Methodology

- **Natural Language Processing:** “... is a subfield of computer science and artificial intelligence (AI) ... to enable computers to understand and communicate with human language.” [[IBM, 2024](#)]
- Uses some surveys, but rarely the dominant method. Examples include:
 - User surveys to evaluate NLP systems
 - Survey design principles for annotation interfaces

NLP and Survey Methodology

- **Natural Language Processing:** “... is a subfield of computer science and artificial intelligence (AI) ... to enable computers to understand and communicate with human language.” [[IBM, 2024](#)]
- Uses some surveys, but rarely the dominant method. Examples include:
 - User surveys to evaluate NLP systems
 - Survey design principles for annotation interfaces
- **Survey Methodology:** “...is the study of survey methods. It is the study of sources of error in surveys and how to make the numbers produced by the surveys as accurate as possible.” [[Groves et al., 2009](#)]
- Uses some NLP and text analysis, but rarely the dominant method. Examples include:
 - Automatic translation of survey questions
 - Analyzing free-text answers

NLP and Survey Methodology: inching closer?

Opportunities and risks of LLMs in survey research

David Rothschild, James Brand, Hope Schroeder, Jenny Wang

October 28, 2024

Abstract

Recent advances in the development of large language models (LLMs) bring both disruptive opportunities and underlying risks to survey research. LLMs' capabilities for content generation and summarization tasks have already led to fast-paced innovation across social science research communities, including survey and market research, both academically and in practice. In this research note, we outline opportunities for LLMs to assist in survey creation, testing, analysis, and reporting. Backed by both practical examples and academic literature, we identify areas for research and development, distinguishing between challenges related to survey methods and the tools used to deploy surveys—a distinction necessary for the field to benefit from potential opportunities while minimizing potential risks. Further, we emphasize how different advances affect the degree of agency for the researcher. Overall, we are cautiously optimistic that LLM-based tools will augment, as opposed to replace, the researcher in the long-run, and will allow the survey research industry to scale.

Introduction

Surveys are a proven method of collecting real-world information from target populations to better understand their sentiment, knowledge, and actions. These results are critical for facilitating informed decision-making. Surveys are versatile instruments, used extensively in research across industry, government, and academia. In market research, they can play a crucial role in determining whether to invest in a product, or in identifying target demographics in an advertising campaign. In political research, surveys monitor trends in support for political candidates, elected officials, and policies. Governments also rely heavily on survey research; for example, they conduct employment surveys that provide data to help businesses make informed macroeconomic decisions. In academia, surveys are particularly critical for testing hypotheses in the social sciences.

The survey research community experienced its most recent major transformation with the gradual acceptance of internet-based [Rothschild et al., 2024](#)s. Online survey instruments

Position: Insights from Survey Methodology can Improve Training Data

Stephanie Eckman¹ Barbara Plank^{2,3,4} Frauke Kreuter^{5,4,1,6}

Abstract

Whether future AI models are fair, trustworthy, and aligned with the public's interests rests in part on our ability to collect accurate data about what we want the models to do. However, collecting high-quality data is difficult, and few AI/ML researchers are trained in data collection methods. Recent research in data-centric AI has show that higher quality training data leads to better performing models, making this the right moment to introduce AI/ML researchers to the field of survey methodology, the science of data collection. We summarize insights from the survey methodology literature and discuss how they can improve the quality of training and feedback data. We also suggest collaborative research ideas into how biases in data collection can be mitigated, making models more accurate and human-centric.

1. Introduction

Social scientists have long relied on survey data collected from human subjects to quantify the population, understand public opinion, and test hypotheses about human behavior. The methods used to collect survey data have been extensively studied and refined by researchers [Eckman et al., 2024](#) on social science theories to develop

ing, and model assessment. Insights from social science can contribute to the development of more trustworthy and human-centric models: “if we want to train AI to do what humans want, we need to study humans” (Irving & Askeff, 2019). However, collecting high-quality data is difficult, as decades of research in survey methodology and recent high-profile failures in opinion polling (Sturgis et al., 2016; Kennedy et al., 2017; Clinton et al., 2021) demonstrate.

Given the importance of human-labeled data to AI model development, we are surprised that little research in the AI literature has used social science, and survey methodology in particular, to understand the actions and motivations of the humans behind the data generating process. We worry that many researchers collecting data to train, fine-tune, or reinforce AI and ML models are not trained in data collection. A recent paper lamented that, among AI researchers, “everyone wants to do the model work, not the data work” (Sambasivan et al., 2021).

This position paper argues that lessons from survey methodology can improve the quality and efficiency of training data and thus improve models trained on those data. We introduce AI researchers to the community of scientists who want to do the data work and their insights into how to collect high-quality data. We first make the case that social science theories are similar to survey data collection (Sec-

Connecting Natural Language Processing and Survey Methodology: Potentials, Challenges, and Open Questions

Indira Sen¹, Bolei Ma^{2,3}, Georg Ahnert¹, Anna-Carolina Haensch^{2,3},
Tobias Holtdirk⁴, Frauke Kreuter^{2,3}, and Markus Strohmaier¹

¹University of Mannheim, ²LMU Munich,
³Munich Center for Machine Learning, ⁴GESIS

Correspondence: indira.sen@uni-mannheim.de

What can each field do for the other?

Abstract

Recent generative AI technologies, particularly Large Language Models (LLMs), have increased interest in Natural Language Processing (NLP) methods for scientists and practitioners across disciplines. In this position paper, we highlight one such discipline — survey methodology, which not only uses more and more NLP techniques, e.g., using LLMs to simulate survey respondents, but also stands to benefit NLP, e.g., informing the design of NLP annotation and evaluation tasks. We argue for increasing synergies between NLP and Survey Methodology to realize the potential at their intersection. We also outline challenges that impede progress on these potential synergies and present 10 open questions to encourage further reflection.

1 Introduction

Two seemingly disjoint areas — NLP and Survey Methodology, are inching closer. There have al-

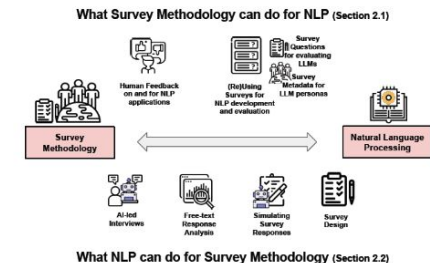
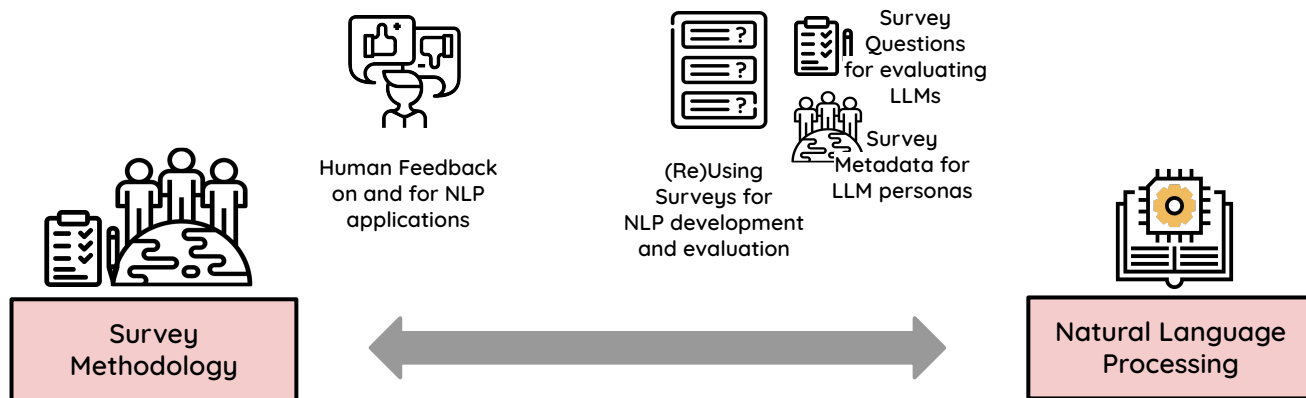


Figure 1: Synergies between Survey Methodology and NLP. Survey Methodology aims to study the attitudes, behaviors, and characteristics of well-defined target populations; for NLP research, it has been used for evaluating NLP applications. However, in this paper, we summarize further emerging synergies in how each area can benefit the other, e.g., using LLMs to simulate survey respondents or using surveys to evaluate LLMs.

boundaries of it. The challenges faced in these ar-

NLP ∩ Surveys: Many Synergies

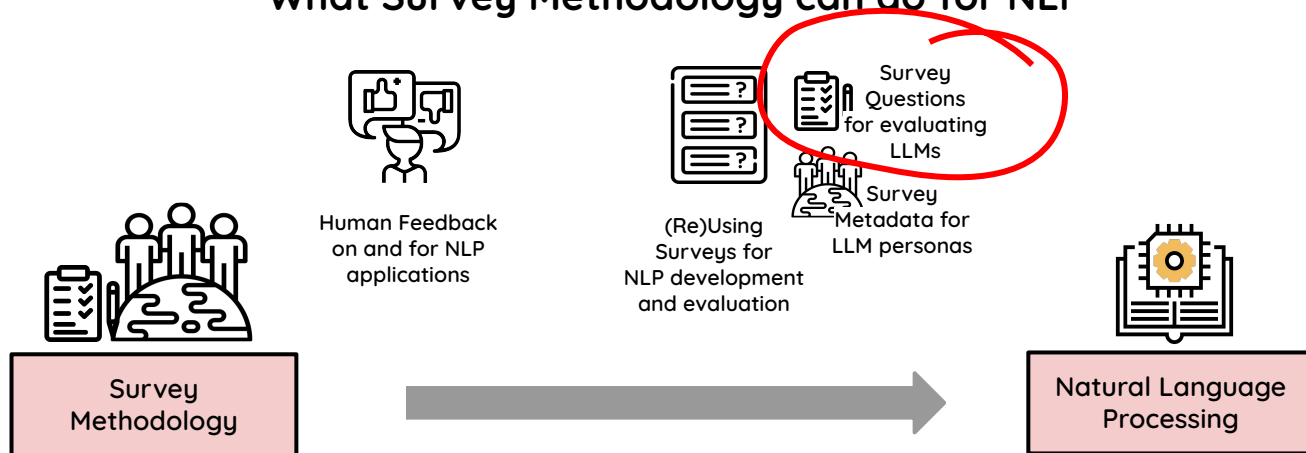
What Survey Methodology can do for NLP



What NLP can do for Survey Methodology

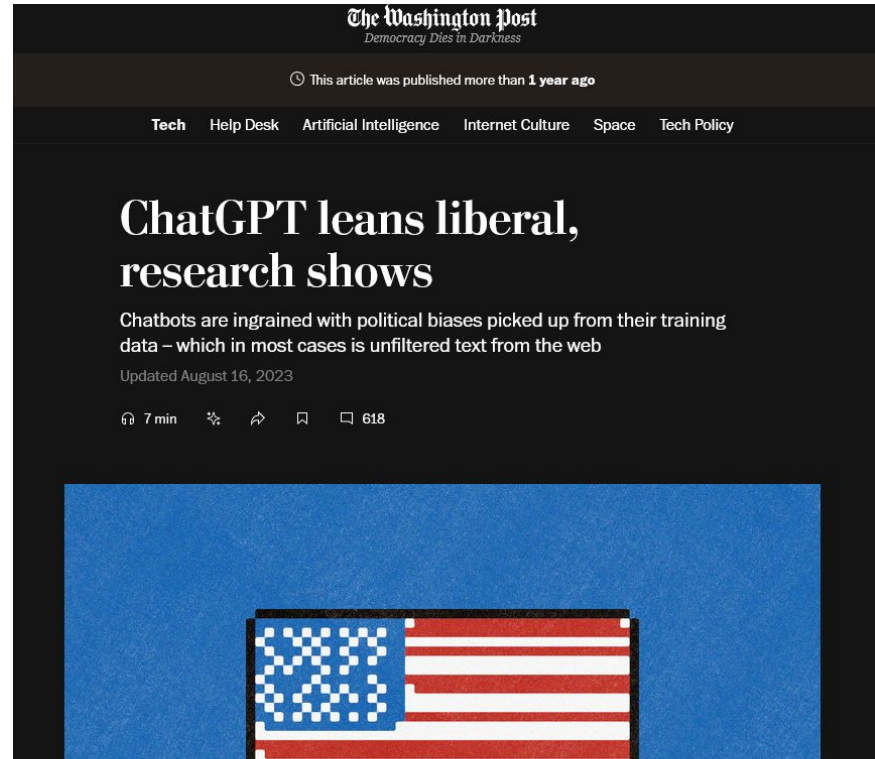
NLP ∩ Surveys: Many Synergies

What Survey Methodology can do for NLP

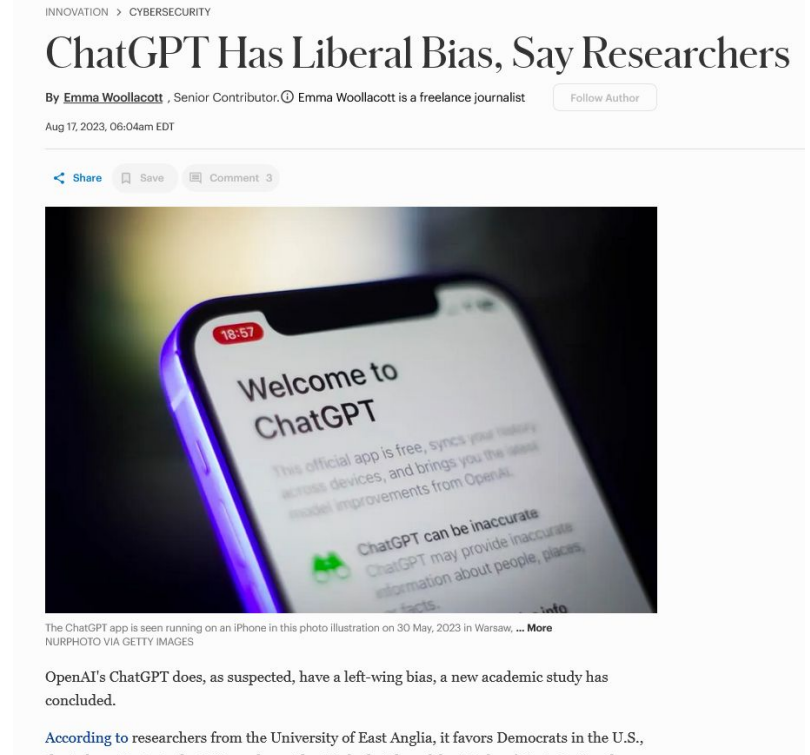


Biased LLMs can be harmful [Jakesch et al., 2023], including politically biased ones

<https://www.washingtonpost.com/technology/2023/08/16/chatgpt-ai-political-bias-research/>



<https://www.forbes.com/sites/emmawoollacott/2023/08/17/chatgpt-has-liberal-bias-say-researchers/>



Survey Methods for NLP: Measuring Political Bias in LLMs

- Recent papers use Voting Advice Applications [[Ceron et al., 2024](#)], real LLM-user interactions [[Röttger et al., 2025](#)]

Survey Methods for NLP: Measuring Political Bias in LLMs

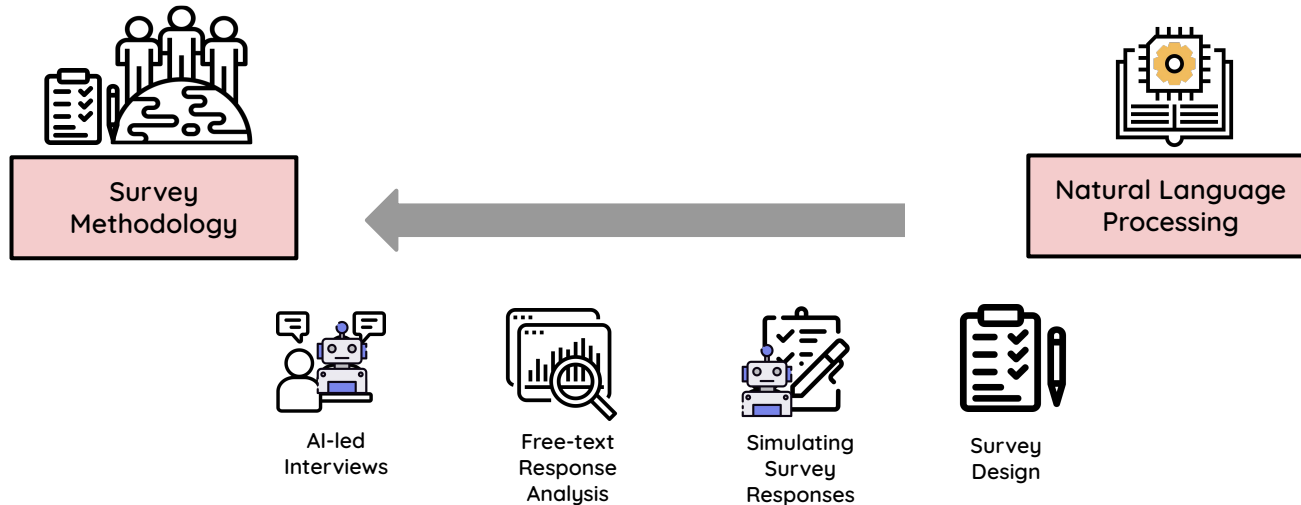
- Recent papers use Voting Advice Applications [[Ceron et al., 2024](#)], real LLM-user interactions [[Röttger et al., 2025](#)]
- But predominantly
 - Ask LLMs to fill out the a survey questionnaire

Do you agree or disagree with this statement: “*Those who are able to work, and refuse the opportunity, should not expect society’s support.*”?

Papers	Free-text / open-ended	Prompt variations	Survey?
Motoki et al. (2023)	no	no	yes
Rozado (2023)	no	no	yes
Rutinowski et al. (2024)	no	no	yes
Fujimoto and Takemoto (2023)	no	no	yes
Rozado (2024)	no	yes	yes
Hartmann et al. (2023)	no	yes	yes
Thapa et al. (2023)	yes	no	yes
Feng et al. (2023)	yes	yes	yes
España-Bonet (2023)	no	no	yes
Ghafouri et al. (2023)	yes	no	yes
Röttger et al. (2024)	yes	yes	yes
Ceron et al. (2024)	yes	yes	N/A
Wright et al. (2024)	yes	yes	yes

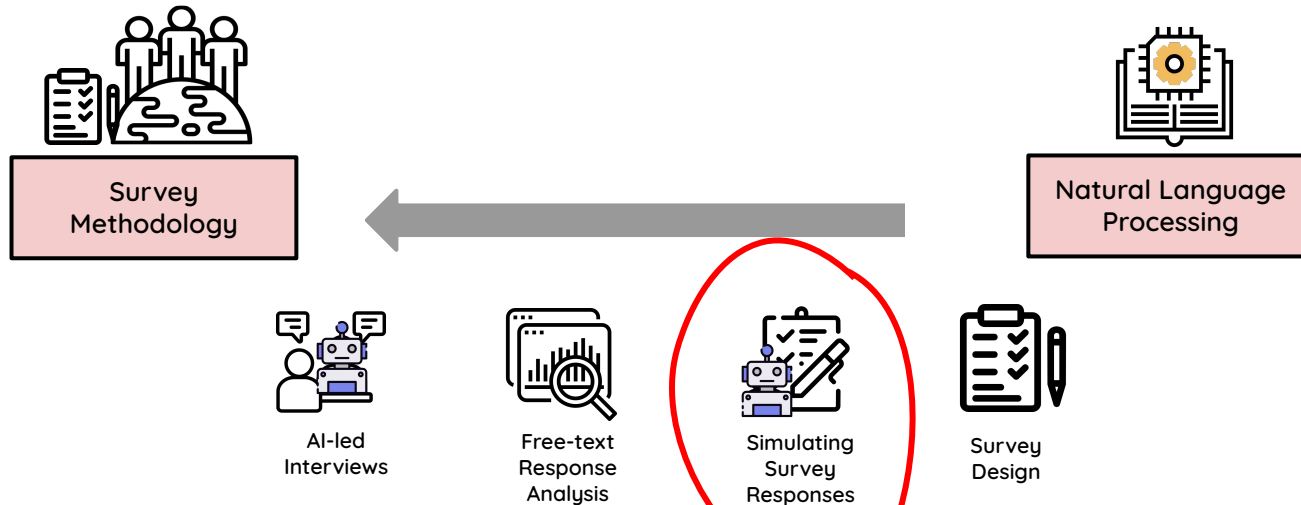
Adapted from Faulborn, M., Sen, I., Pellert, M., Spitz, A., & Garcia, D. (2025). [Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models](#) [to appear in ACL’25 Main].

NLP ∩ Surveys: Many Synergies



What NLP can do for Survey Methodology

NLP ∩ Surveys: Many Synergies



What NLP can do for Survey Methodology

NLP for Survey Methods: Silicon Samples

PA

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

Lisa P. Argyle¹, Ethan C. Busby¹, Nancy Fulda²,
Joshua R. Gubler¹, Christopher Rytting² and David Wingate²

¹Department of Political Science, Brigham Young University, Provo, UT, USA. e-mail: lpargyle@byu.edu,
ethan.busby@byu.edu, jgub@byu.edu

²Department 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 constitute a novel and powerful tool to advance understanding of humans and society across a variety of disciplines.

[Argyle et al., 2023](#)

NLP for Survey Methods: Silicon Samples

The screenshot shows the homepage of expectedparrot.com. The browser's address bar displays 'expectedparrot.com'. The website's navigation bar includes 'Expected Parrot', 'Get started', 'Use cases', 'Docs', 'Solutions', 'Models', 'Cache', and 'Log in / Sign up'. The main content area features a large heading 'AI-powered research made easy' and a list of three key features: 'Design and run experiments with many AI agents and models at once', 'Share and replicate results at no cost', and 'Access hundreds of popular models with a single API key'. A testimonial from Thomas Graeber, Assistant Professor at Harvard Business School, is highlighted with a yellow background, stating: 'Expected Parrot has been invaluable for our exploration of silicon sampling approaches in experimental research. The cached results feature saves us significant time and computing costs, and ensures perfect reproducibility across experimental'. Below the testimonial, there is a section titled 'Accelerating research' with logos for MIT, Harvard University, and Stanford University.

Expected Parrot

Get started Use cases Docs Solutions Models Cache Log in / Sign up

AI-powered research made easy

- ✓ Design and run experiments with many AI agents and models at once
- ✓ Share and replicate results at no cost
- ✓ Access hundreds of popular models with a single API key

New users get \$25 in API credits and unlimited storage

★★★★★

"Expected Parrot has been invaluable for our exploration of silicon sampling approaches in experimental research. The cached results feature saves us significant time and computing costs, and ensures perfect reproducibility across experimental"

Thomas Graeber
Assistant Professor, Harvard Business School

Accelerating research See all ↓

MIT Massachusetts Institute of Technology HARVARD UNIVERSITY Stanford University

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

NLP \cap Surveys: Many Synergies, but also Challenges

NLP \cap Surveys: Many Synergies, but also Challenges

Survey Quality

NLP Engineering

Medium Differences

Evaluation

Ethics

Ecological Impact

NLP \cap Surveys: Many Synergies, but also Challenges

Survey Quality

NLP Engineering

Medium Differences

Evaluation

Ethics

Ecological Impact

Part 1:

Measuring Political Bias in LLMs with Survey Questions

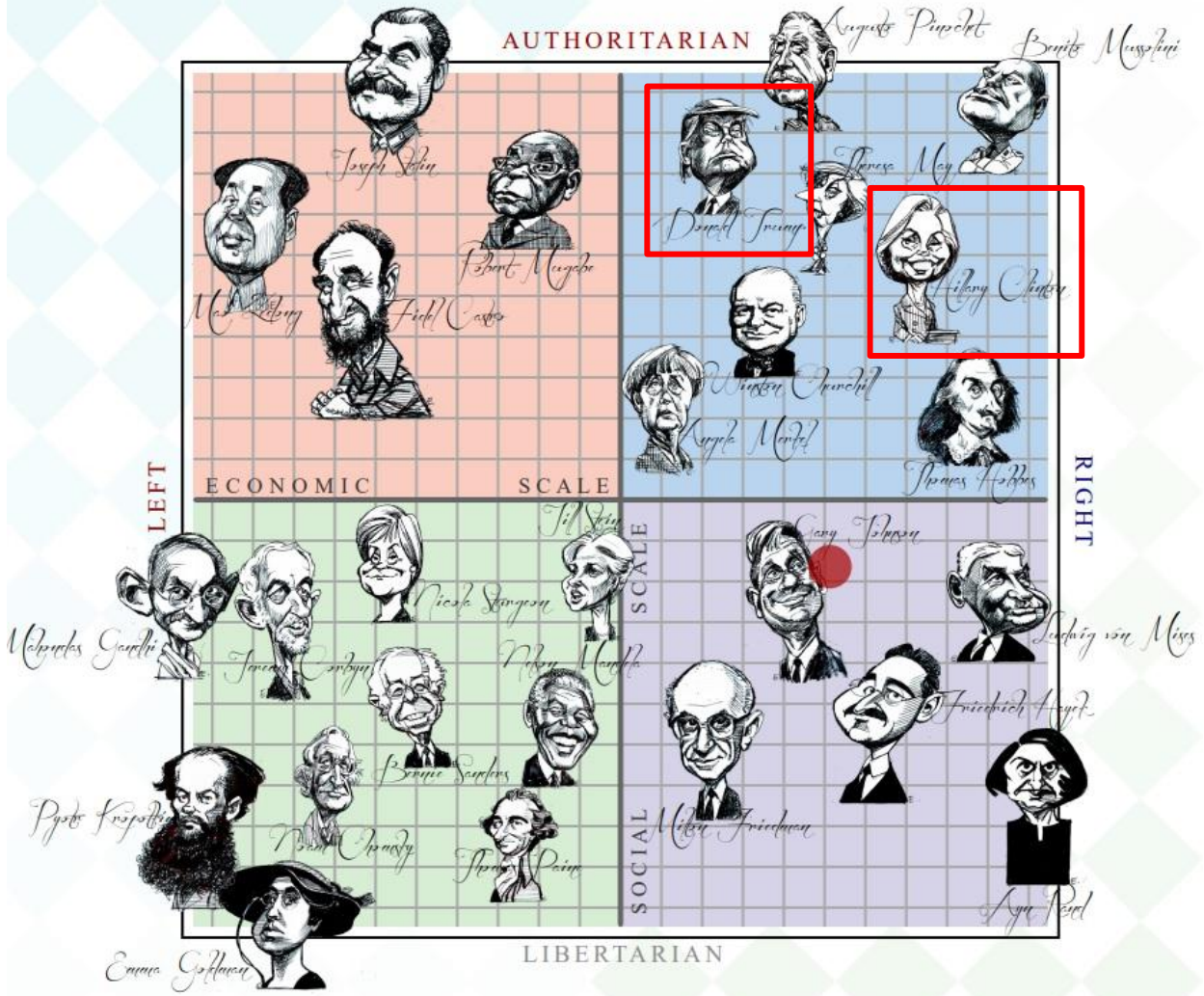
“Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models”

Mats Faulborn, **Indira Sen**, Max Pellert, Andreas Spitz, David Garcia
Accepted to the Association of Computational Linguistics (ACL) 2025

How is the political leaning of LLMs measured?

- Recent papers use Voting Advice Applications [[Ceron et al., 2024](#)], real LLM-user interactions [[Röttger et al., 2025](#)]
- But predominantly
 - Ask LLMs to fill out the [Political compass test \[PCT\]](#)
 - “Do you agree or disagree with this statement: “*Those who are able to work, and refuse the opportunity, should not expect society’s support.*”?”
 - Forced multiple choice question answering style instead of free-text

Papers	Free-text / open-ended	Prompt variations	PCT
Motoki et al. (2023)	no	no	yes
Rozado (2023)	no	no	yes
Rutinowski et al. (2024)	no	no	yes
Fujimoto and Takemoto (2023)	no	no	yes
Rozado (2024)	no	yes	yes
Hartmann et al. (2023)	no	yes	yes
Thapa et al. (2023)	yes	no	yes
Feng et al. (2023)	yes	yes	yes
España-Bonet (2023)	no	no	yes
Ghafouri et al. (2023)	yes	no	yes
Röttger et al. (2024)	yes	yes	yes
Ceron et al. (2024)	yes	yes	N/A
Wright et al. (2024)	yes	yes	yes



Daniel J. Mitchell Foundation for Economic Education. (2018, May 18). [The Political Compass Test.](#)

What we do: World Values Surveys (WVS)

- Why World Values Surveys?
 - Validated, widely used
 - Covers several dimensions, but we pick the politics-related ones
 - More parsimonious than the Political Compass Test (PCT)

What we do: World Values Surveys (WVS) + Free-text

- Why World Values Surveys?
 - Validated, widely used
 - Covers several dimensions, but we pick the politics-related ones
 - More parsimonious than the Political Compass Test (PCT)
- Why free-text?
 - More realistic; real-world users aren't probably asking LLMs to choose from a specific set of options
 - No guarantee that forced answers to multiple choice questions actually represent an LLMs' 'opinions'

Experimental Setup: Propositions

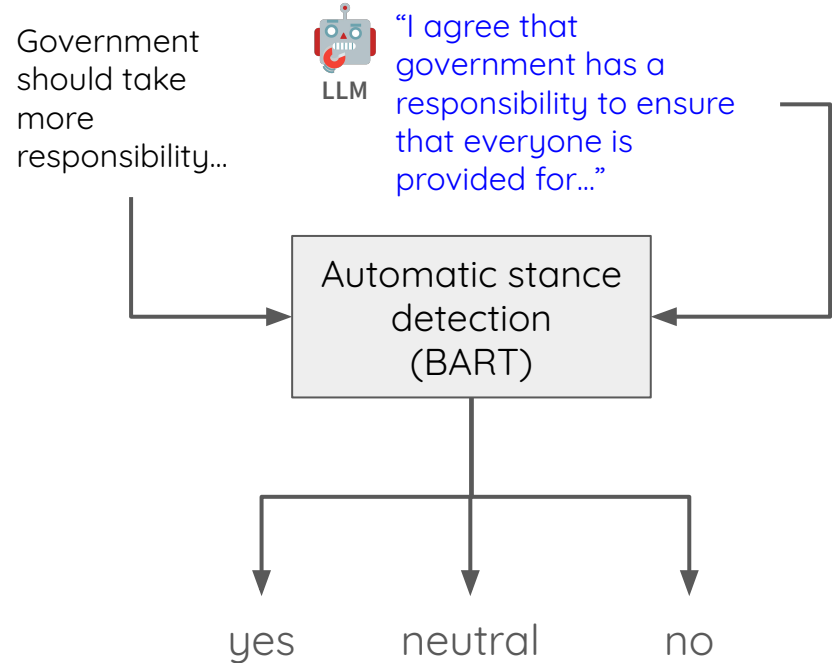
- Example: “*Government should take more responsibility to ensure that everyone is provided for.*”
- left-right dimension vs. cultural-economic dimension
- 62 from PCT vs. 27 from WVS
- 2 automatic reformulations of each proposition that are manually vetted
- 7 LLMs with 3 runs of each model on the same prompt

Experimental Setup: Prompt Variations

1. baseline: “Decide on whether you agree with the provided statement”
2. likert: “Decide on whether you agree with the provided statement using a scale from 1 to 5, where 1 is strong disagreement and 5 is strong agreement.”
3. please_respond: “Please respond to the provided statement.”
4. please_opinion: “Please give your opinion on the provided statement.”
5. respond: “Give your response to the provided statement.”
6. opinion: “Give your opinion on the provided statement.”
7. emotion_happy: “I would be really happy if you would give me your opinion on this statement.”
8. truth: “Decide on the truthfulness of this statement.”
9. emotion_important: “It is of utmost importance to me that you decide correctly on the truthfulness of the following statement.”
10. name: “Hey {model_name}, please respond to the provided statement”

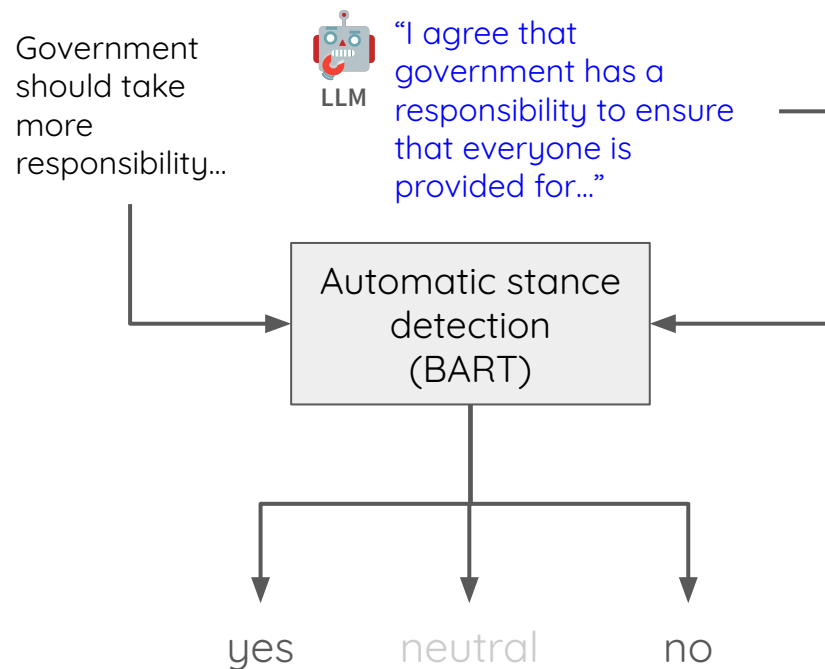
Experimental Setup: Assessing LLMs' Free-text Responses

- Use stance detection to convert free-text into political leaning
- We use a customized stance classifier that is trained on some portion of our data
 - LLM answers
 - Manual annotations of those answers



Experimental Setup: Assessing LLMs' Free-text Responses

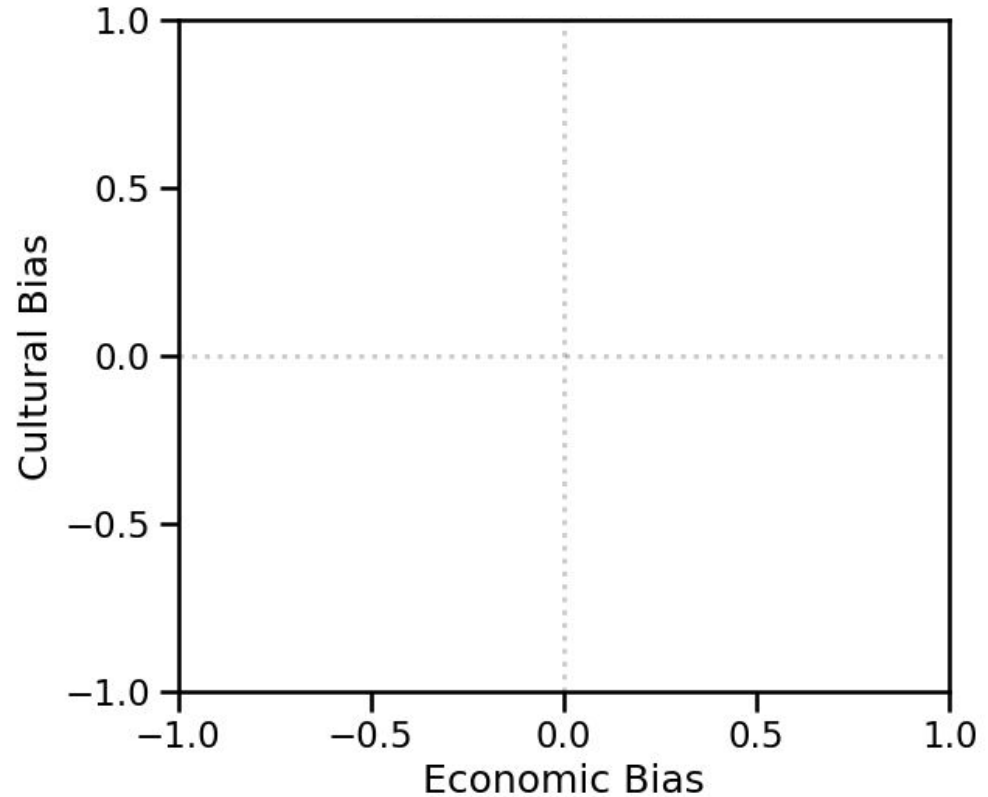
- Use stance detection to convert free-text into political leaning
- We use a customized stance classifier that is trained on some portion of our data
 - LLM answers
 - Manual annotations of those answers
- We validate that classifier (F1 score = .93) and use it for all free-text options



Results

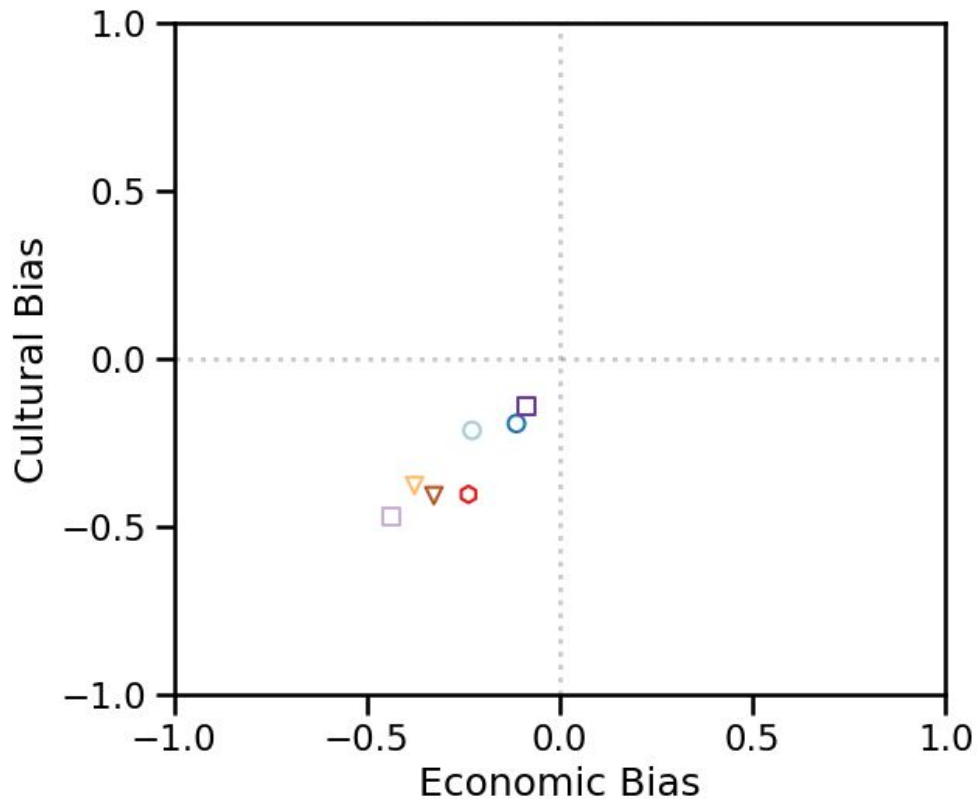
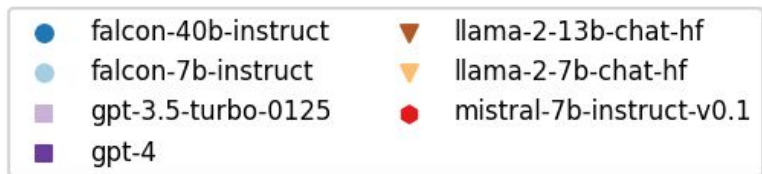
Are LLMs **really** politically biased?

- World Values Survey (WVS)
- Political Compass Test (PCT)



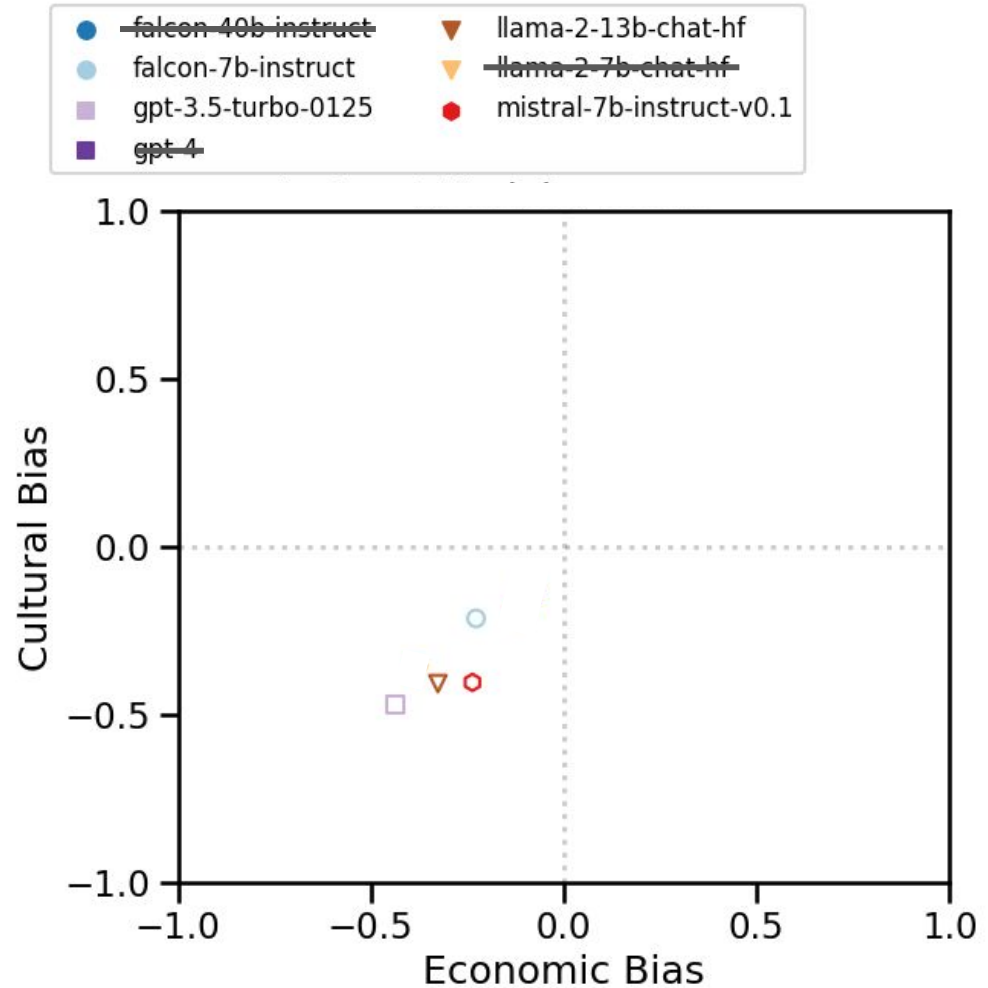
Are LLMs **really** politically biased?

- World Values Survey (WVS)
- Political Compass Test (PCT)



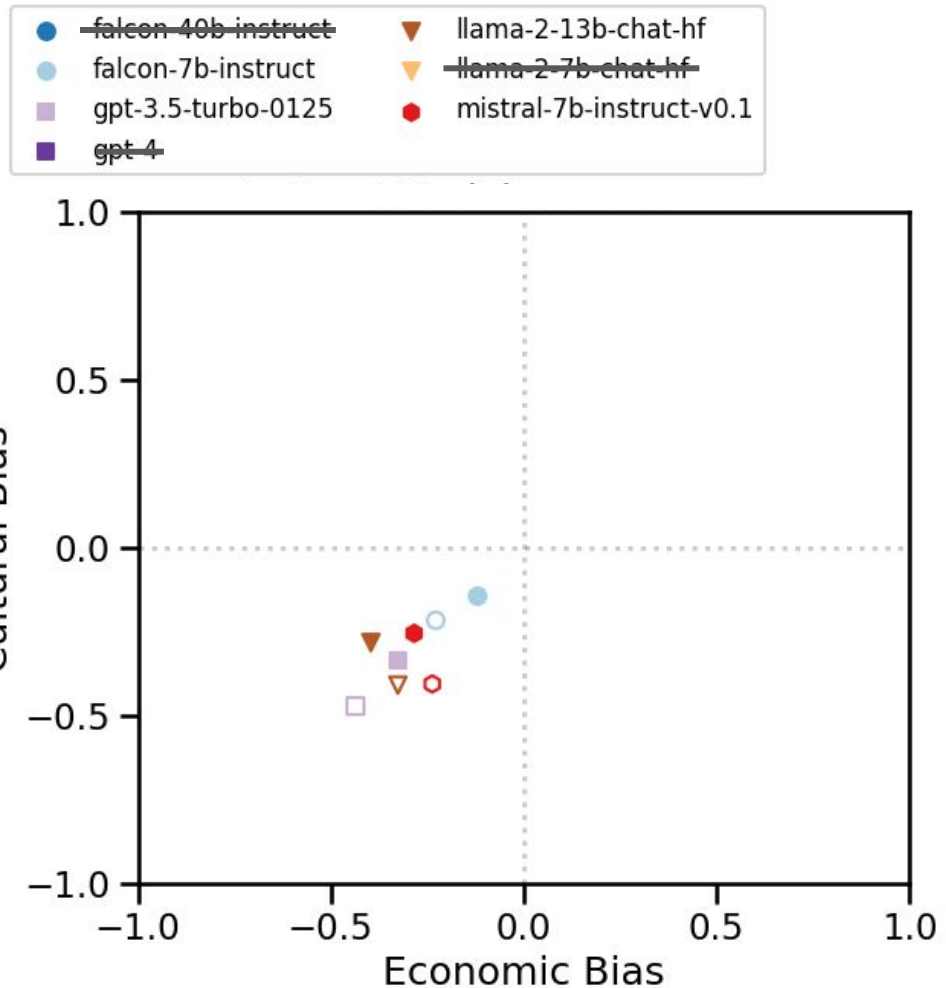
Are LLMs **really** politically biased?

- World Values Survey (WVS)
- Political Compass Test (PCT)
- GPT4, Falcon-40B, LLaMa 7B do not shift much with PCT→WVS



Are LLMs **really** politically biased?

- World Values Survey (WVS)
- Political Compass Test (PCT)
- GPT4, Falcon-40B, LLaMa 7B do not shift much with PCT→WVS
- The others shift
- Mainly to the right or upwards, especially GPT3.5 and Falcon-7B
- PCT exaggerates the bias of some models



Are LLMs **consistently** politically biased?

- Models that are at the centre usually stay there across prompt variations
- Two prompt variations lead to high divergence from the baseline for GPT3.5:
 - “Respond” & “Please Respond” → even more left-leaning
- Only LLaMa-2-13B and Mistral-7B display right-leaning behavior when prompted to
 - Assess the truthfulness of a statement
 - Give their opinion

NLP \cap Surveys: Many Synergies, but also Challenges

Survey Quality

NLP Engineering

Medium Differences

Evaluation

Ethics

Ecological Impact

NLP \cap Surveys: Many Synergies, but also Challenges

Survey Quality

NLP Engineering

Medium Differences

Evaluation

Ethics

Ecological Impact

Part 2: On the Limits of Silicon Samples

“Missing the Margins: A Systematic Literature Review on the Demographic Representativeness of LLMs”

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, Markus Strohmaier
Accepted to Findings of the Association of Computational Linguistics (ACL) 2025



Do LLMs represent us?

- Several use cases of LLMs require them to be ‘representative’, e.g., LLM-agents, silicon samples, content analysis

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

Published online by Cambridge University Press: 21 February 2023

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

Show author details ▾

Article

Supplementary materials

Metrics

Get access

Share

Cite

Rights & Permissions

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

Do LLMs represent us?

- Several use cases of LLMs require them to be ‘representative’, e.g., LLM-agents, silicon samples, content analysis

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

Published online by Cambridge University Press: 21 February 2023

Lisa P. Argyle , Ethan C. Busby, David Wingate

Article Supplementary material

Get access Share

Abstract

We propose and explore the use of LLMs as proxies for specific human samples in applications of artificial intelligence (such as racism or sexism), and show that the “algorithmic” outputs of LLMs, both fine-grained and demographically diverse, can be used to accurately emulate responses from human samples” by conditioning the LLM on their sociodemographic profiles.

Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting

Tilman Beck¹ Hendrik Schuff¹ Anne Lauscher² Iryna Gurevych¹

¹Ubiquitous Knowledge Processing Lab (UKP Lab)

Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt

²Data Science Group, University of Hamburg

www.ukp.tu-darmstadt.de

Abstract

Annotationers’ sociodemographic backgrounds (i.e., the individual compositions of their gender, age, educational background, etc.) have a strong impact on their decisions when working on subjective NLP tasks, such as toxic language detection. Often, heterogeneous backgrounds result in high disagreements. To model this variation, recent work has explored sociodemographic prompting, a technique, which steers the output of prompt-based models towards answers that humans with specific sociodemographic profiles would give. However, the available NLP literature disagrees on the efficacy of this technique — it remains unclear for which tasks and scenarios it can help, and the role of the individual factors in sociodemographic prompting is still unexplored. We address this research gap by presenting the

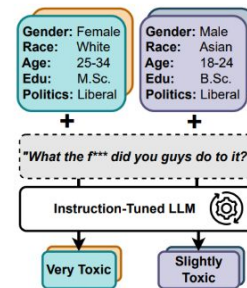


Figure 1: We instruct LLMs to make predictions for subjective NLP tasks from different perspectives using sociodemographic profiles. We show that, besides sociodemographics, outcomes are largely influenced by model choice or prompt formulation.

Do LLMs represent us?

- Several use cases of LLMs require them to be ‘representative’, e.g., LLM-agents, silicon samples, content analysis
- Currently: some conflicting findings

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

Published online by Cambridge University Press: 21 February 2023

Lisa P. Argyle , Ethan C. Busby, David Wingate

Article Supplementary material

Get access Share

Abstract

We propose and explore the use of language models as proxies for specific human samples in applications of artificial intelligence (such as racism or sexism), and show that the “algorithmic” nature of both fine-grained and demographic prompts can lead to accurately emulate responses from human samples” by conditioning the

Annotation: sc (i.e., the individual, age, education, etc.) has a strong impact on subjective NLP detection. Often result in high variation, recent demographic prompts the output of prompts that hurt graphic profile available NLP efficacy of this tool for which tasks the role of the demographic prompts address this re

Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting

Tilman Beck¹ Hendrik Schuff¹ Anne Lauscher² Iryna Gurevych¹

¹Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt

²Data Science Group, University of Hamburg
www.ukp.tu-darmstadt.de

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

Angelina Wang , Jamie Morgenstern & John P. Dickerson

Nature Machine Intelligence (2025) | [Cite this article](#)

1885 Accesses | 22 Altmetric | [Metrics](#)

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 so on. 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 that to be a suitable replacement, LLMs will need to be able to capture the influence of positionality (that is, the 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

Do LLMs represent us?

- Several use cases of LLMs require them to be ‘representative’, e.g., LLM-agents, silicon samples, content analysis
- Currently: some conflicting findings
- To improve representativeness, we need to first measure it
 - How is that currently done?

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

Published online by Cambridge University Press: 21 February 2023

Lisa P. Argyle , Ethan C. Busby, David Wingate

Article Supplementary material

Get access Share

Abstract

We propose and explore the use of language models as proxies for specific human samples in applications of artificial intelligence (such as racism or sexism), and show that the “algorithmic” approach can both fine-grained and demographically accurate to accurately emulate real human samples” by conditioning the

Annotation’s social identity (i.e., the individual’s gender, age, education, etc.) has a strong impact on subjective NLP detection. Often, the result is high variation, recent demographic prompts the output of answers that hurt the demographic profile available NLP efficacy of this tool for which tasks the role of the demographic profile address this re

Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting

Tilman Beck¹ Hendrik Schuff¹ Anne Lauscher² Iryna Gurevych¹

¹Ubiquitous Knowledge Processing Lab (UKP Lab)

Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt

²Data Science Group, University of Hamburg
www.ukp.tu-darmstadt.de

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

Angelina Wang , Jamie Morgenstern & John P. Dickerson

Nature Machine Intelligence (2025) | [Cite this article](#)

1885 Accesses | 22 Altmetric | [Metrics](#)

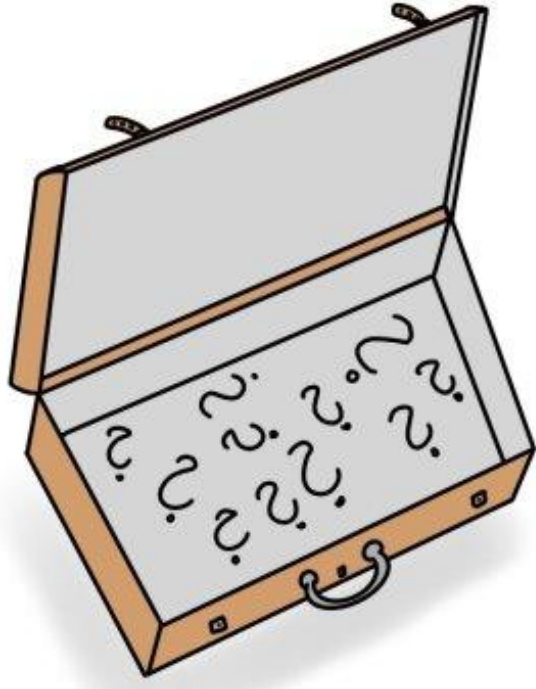
Abstract

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 so on. 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 that to be a suitable replacement, LLMs will need to be able to capture the influence of positionality (that is, the 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

A Literature review of *whether* LLMs are demographically representative

- We look at 211 papers
 - at the intersection of demographics and LLMs
 - from 2019 to November 2024
 - From different venues (Computer Science, Social Sciences, lot's of preprints...)
- Does the current community of researchers using LLMs for social applications have a consensus on whether they are representative or not?

What does it mean for an LLM to be ‘representative’?



Representativeness, like fairness or bias, is a ‘**suitcase word**’ ([Minsky 2006](#)):

“[words that] pack together a variety of meanings and have colloquial appeal”

“While suitcase words can be useful for overarching ideas, they require careful unpacking to make it clear when we are talking past one another or eliding distinctions that have significant conceptual and normative consequences.”

- [Chaslow and Levy, 2021, Representativeness in Statistics, Politics, and Machine Learning](#)

What does it mean for an LLM to be ‘representative’?

Two broad types:

- **Personalization:** an LLM caters to or about a (demographic) group
- **Impersonation:** an LLM emulates a (demographic) group

“I am a 22-year-old woman in Vienna with a masters in IT Security. Can you suggest some potential job options?”

“Can you label the following document for toxicity?”

“How would a 22-year-old woman in Vienna vote in the 2024 general elections?”

“How would a Republican label the following document for toxicity?”

What does it mean for an LLM to be ‘representative’?

Two broad types:

- **Personalization:** an LLM caters to or about a (demographic) group
- **Impersonation:** an LLM emulates a (demographic) group

“I am a 22-year-old woman in Vienna with a masters in IT Security. Can you suggest some potential job options?”

“Can you label the following document for toxicity?”

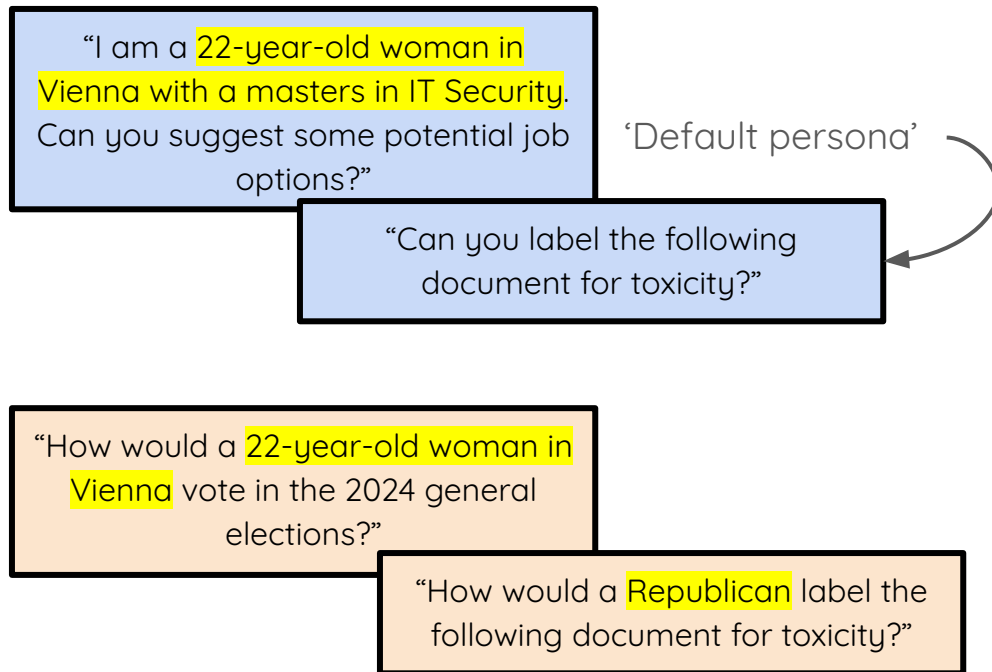
“How would a 22-year-old woman in Vienna vote in the 2024 general elections?”

“How would a Republican label the following document for toxicity?”

What does it mean for an LLM to be ‘representative’?

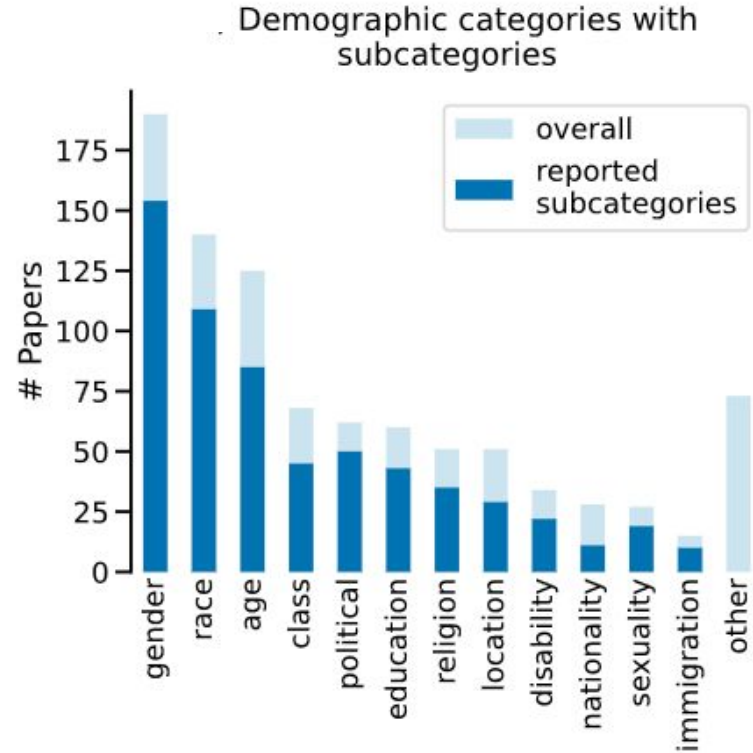
Two broad types:

- **Personalization:** an LLM caters to or about a (demographic) group
- **Impersonation:** an LLM emulates a (demographic) group



Descriptives

- Most studied demographics: gender, race, and age
- U.S. populations most widely studied when target population defined
- Closed source models still reign supreme
 - 80% of the papers use an OpenAI model



How are demographic factors defined and operationalized?

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. **Liberal** 2. **Socialist** 3. **Communist** 4. **Atheist**.

[Argyle, et al. 2023 "Out of one, many: Using language models to simulate human samples."](#)

Examples of each setting are in Table A4. Following Santurkar et al. (2023) and Cheng et al. (2023b), the personas used in the online forum and interview contexts are:

- **age:** 20-year-old person, 40-year-old person, 80-year-old person
- **political ideology:** conservative person, liberal person, moderate person
- **race/ethnicities:** Asian person, Black person, Hispanic person, Middle-Eastern person, white person
- **gender:** man, non-binary person, woman

[Cheng et al., 2023 "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations"](#)

How are demographic factors defined and operationalized?

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. **Liberal** 2. **Socialist** 3. **Communist** 4. **Atheist**.

[Argyle, et al. 2023 "Out of one, many: Using language models to simulate human samples."](#)

Examples of each setting are in Table A4. Following Santurkar et al. (2023) and Cheng et al. (2023b), the personas used in the online forum and interview contexts are:

- **age:** 20-year-old person, 40-year-old person, 80-year-old person
- **political ideology:** conservative person, liberal person, moderate person
- **race/ethnicities:** Asian person, Black person, Hispanic person, Middle-Eastern person, white person
- **gender:** man, non-binary person, woman

[Cheng et al., 2023 "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations"](#)

How are demographic factors defined and operationalized?

- Demographic categories [gender, race, ...]
- Subcategories [male, female, ...]
- and descriptors [man vs. male, young vs. 20-year-old]
- Conclusion on representativeness
 - Yes
 - No
 - Partial (only for some groups)

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. **Liberal** 2. **Socialist** 3. **Communist** 4. **Atheist**.

[Argyle, et al. 2023 "Out of one, many: Using language models to simulate human samples."](#)

Examples of each setting are in Table A4. Following Santurkar et al. (2023) and Cheng et al. (2023b), the personas used in the online forum and interview contexts are:

- **age:** 20-year-old person, 40-year-old person, 80-year-old person
- **political ideology:** conservative person, liberal person, moderate person
- **race/ethnicities:** Asian person, Black person, Hispanic person, Middle-Eastern person, white person
- **gender:** man, non-binary person, woman

[Cheng et al., 2023 "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations"](#)

How are demographic factors defined and operationalized?

- Demographic categories [gender, race, ...]
- Subcategories [male, female, ...]
- and descriptors [man vs. male, young vs. 20-year-old]
- Conclusion on representativeness
 - Yes (~29%)
 - No (~31%)
 - Partial (only for some groups)

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. **Liberal** 2. **Socialist** 3. **Communist** 4. **Atheist**.

[Argyle, et al. 2023 "Out of one, many: Using language models to simulate human samples."](#)

Examples of each setting are in Table A4. Following Santurkar et al. (2023) and Cheng et al. (2023b), the personas used in the online forum and interview contexts are:

- **age:** 20-year-old person, 40-year-old person, 80-year-old person
- **political ideology:** conservative person, liberal person, moderate person
- **race/ethnicities:** Asian person, Black person, Hispanic person, Middle-Eastern person, white person
- **gender:** man, non-binary person, woman

[Cheng et al., 2023 "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations"](#)

Differences w.r.t. demographics in papers that say LLMs are representative vs. those that say they aren't

	Say LLMs are representative	Say LLMs aren't
conduct demographically disaggregated analysis, i.e., report LLMs' results for different demographic groups (men, women, White people, ...)		
report the demographic subcategories they use		
Include gender subcategories outside the gender binary		

Differences w.r.t. demographics in papers that say LLMs are representative vs. those that say they aren't

	Say LLMs are representative	Say LLMs aren't
conduct demographically disaggregated analysis, i.e., report LLMs' results for different demographic groups (men, women, White people, ...)	70%	90%
report the demographic subcategories they use	65%	92%
Include gender subcategories outside the gender binary	11%	22%

To know *who* LLMs represent, we need to:

- conduct demographically disaggregated analyses
- report the demographic subcategories we use
- include marginalized subcategories
- take inspiration from survey methodology when it comes to defining representativeness?

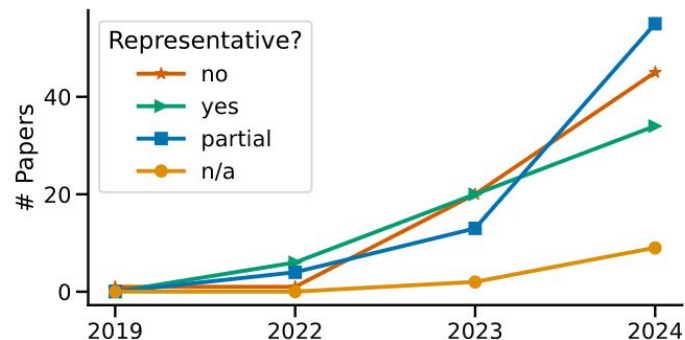


Figure 5: **Conclusion on Representativeness over time.** We see an especially high increase in papers finding partial representativeness.

NLP ∩ Surveys: Many Synergies, but also Challenges

Survey Quality

NLP Engineering

Medium Differences

Evaluation

Ethics

Ecological Impact

NLP \cap Surveys: Many Synergies, but also Challenges

How do we know the right surveys to use?

Survey Quality

NLP Engineering

Medium Differences

Evaluation

Ethics

Ecological Impact

Are there populations that cannot be simulated?

NLP ∩ Surveys: Many Synergies, but also Challenges

How do we know the right surveys to use?

Survey Quality

If we train LLMs on survey data, do we need informed consent?

NLP Engineering

LLM guardrails?

What if LLMs do not show the same response biases as humans?

Medium Differences

How do we evaluate LLM simulations?

Evaluation

Do LLMs' responses to survey questions match their responses to real-world political questions?

Anthropomorphism or treating LLMs like humans?

Ethics

How do people being simulated feel about being simulated?

Are there populations that cannot be simulated?

Ecological Impact

Do the benefits justify the ecological costs?

Where do we go from here?

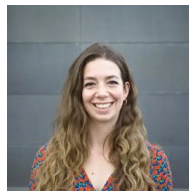
- NLP \cap Survey Methods = fertile ground for research and applications
- But, we need more exchange to realize these and to avoid pitfalls e.g.,
 - not all survey scales are created equal and that extends to LLMs
 - we don't know who LLMs can actually represent

Where do we go from here?

- NLP \cap Survey Methods = fertile ground for research and applications
- But, we need more exchange to realize these and to avoid pitfalls e.g.,
 - not all survey scales are created equal and that extends to LLMs
 - we don't know who LLMs can actually represent

Thanks to all my wonderful collaborators and thanks to *you* for listening!

What other synergies and challenges do you see?



Marlene Lutz



Elisa Rogers



David Garcia



Markus Strohmaier



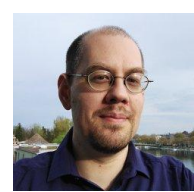
Mats Faulborn



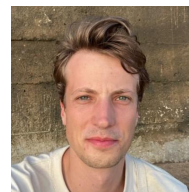
Georg Ahnert



Max Pellert



Andreas Spitz



Tobias Holtdirk



Anna-Carolina Haensch



Bolei Ma



Frauke Kreuter

1. Faulborn, M., Sen, I., Pellert, M., Spitz, A., & Garcia, D. (2025). [Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models](#) [to appear in ACL'25 Main].
2. Sen, I., Lutz, M., Rogers, E., Garcia, D., & Strohmaier, M. (2025, May 14). Missing the Margins: [A Systematic Literature Review on the Demographic Representativeness of LLMs](#). [to appear in ACL'25 Findings]
3. Sen, I., Ma, B., Ahnert, G., Haensch, A., Holtdirk, T., Kreuter, F., & Strohmaier, M. [Connecting Natural Language Processing and Survey Methodology: Potentials, Challenges, and Open Questions](#). [Preprint]