Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
September 2023

Whose Opinions Do Language Models Reflect?

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, Tatsunori Hashimoto

SINCE THE NOVEMBER 2022 DEBUT OF CHATGPT, an AI chatbot developed by OpenAI, language models have been all over the news. But as people use chatbots—to write stories and look up recipes, to make travel plans and even further a real estate business—journalists, policymakers, and members of the public are increasingly paying attention to the important question of whose opinions these language models reflect. In particular, one emerging concern is that AI-generated text may be able to influence our views, including political beliefs, without our realizing it.

Language models, chatbots included, are shaped by a variety of data inputs. These inputs are provided by internet users in the form of training data (such as the authors of internet comments or blogs), crowd workers providing feedback on how to improve data or models (as OpenAI used in Kenya), and the developers themselves (who make high-level decisions regarding data collection and training). The data used to inform language models therefore represents a range of individuals and draws on a wide variety of opinions about sports, politics, culture, food, and many other topics. Meanwhile, language models are being asked subjective questions that have no clear right or wrong answer.

In our paper, "Whose Opinions Do Language Models Reflect?," we introduce a quantitative framework to answer this very question. The framework includes our development of a dataset to evaluate language models' alignment with 60 demographic groups in the United States, covering a diversity of topics. Using

Key Takeaways

Language models are shaped by a variety of inputs and opinions, from the individuals whose views are included in the training data to crowd workers who manually filter that data.

We found that language models fine-tuned with human feedback—meaning models that went through additional training with human input—were less representative of the general public's opinions than models that were not fine-tuned.

It is possible to steer a language model toward the opinions of a particular demographic group by asking the model to respond as if it were a member of that group, but this can lead to undesirable side effects, such as exacerbating polarization and creating echo chambers.

We highlight the need for further research on the evaluation of language models that can help policymakers and regulators quantitatively assess language model behavior and compare it to human preferences and opinions.

this framework, we find a major gap between the responses provided by language models and the views of demographic groups in the United States. We also discover a number of U.S. groups whose views are poorly reflected by current language models, such as people 65 years of age and older, widowed individuals, and people who regularly attend religious services. Our framework allows policymakers to quantitatively evaluate language models and serves as a reminder that issues of representation in language models should remain front of mind.

> *Language models are being asked subjective questions that have no clear right or wrong answer.*

Introduction

Previous work has examined subcomponents of the question of what views language models are shaped by—and whose opinions those views reflect. Some studies have found that on certain topics, such as gun rights, earlier language model versions (such as ChatGPT 3.5) express views typically associated with left-leaning individuals. Since then, OpenAI and other chatbot developers have been trying to fix political biases in the subsequent versions of their models, including exploring how they may be trained to generate different perspectives and worldviews. Other studies have found that language models, with the right training, can mimic certain demographic groups' tendencies. For example, they can be conditioned to support the views of a presidential candidate for whom certain people might vote. These are significant findings for researchers, the users of language models, and policymakers.

In our paper, we rely on public opinion surveys to study the behavior of nine language models in their earlier versions (such as the GPT-3 model family). It is possible to test language models' responses to issues in an ad hoc fashion—for example, by coming up with lists of topics and then generating questions around those topics. But public opinion surveys offer several advantages: Experts choose the survey topics, they work to word the questions unambiguously and with nuance, and they make each question multiple choice, which makes it easily adaptable for a language model prompt. Despite their limits, public opinion surveys are already used across the United States to survey people's views on politics, consumer behavior, and other subjects.

With this in mind, we built a dataset of 1,498 questions from 15 polls conducted by Pew Research's American Trends Panel. Each poll had thousands of U.S. respondents, and the questions covered topics spanning politics, health, privacy, personal relationships, and other areas. Then, we fed nine language models those questions and evaluated the responses across three axes: representativeness, steerability, and consistency. These axes correspond to particular questions about whose opinions language models reflect:

- Representativeness: How aligned is the default language model response distribution with the general U.S. population (or a demographic subgroup)?

- Steerability: Can a language model emulate the opinion distribution of a group (such as Democrats or Republicans) with additional prompting?

- Consistency: Are the groups that language models align with consistent across topics?

> *Despite their limits, public opinion surveys are already used across the United States to survey people's views on politics, consumer behavior, and other subjects.*

Research Outcomes

We developed a formula for quantifying how much the distribution of opinions generated by language models on a given topic differs from the distribution of opinions expressed by humans in response to a Pew survey on that topic. Our findings reveal that a randomly selected demographic group's opinions are more representative of the general public than the views expressed by the language models. The opinions expressed by most language models are about as aligned with the overall populace as those of agnostic and orthodox individuals on abortion or Democrats and Republicans on climate change. In addition, language models are particularly unrepresentative of several groups, including people aged 65 and older, widowed individuals, and people who regularly attend religious services.

Interestingly, our analysis revealed that of the language models we examined, those fine-tuned with human feedback—meaning those that underwent additional training with human input—were less representative of the opinions of the general public than models that were not fine-tuned. Particularly, language models tuned with reinforcement learning from human feedback (RLHF)—a training technique that rewards models for mimicking human responses often collected from crowd workers and amplifying the perspectives that lead to higher rewards—are more aligned with left-leaning, liberal views.

To adjust for these issues within those specific groups, we attempted to steer language models toward one of multiple demographic groups, such as Republicans or Asians, by prompting them to behave like these groups. Most of the language models we steered in this fashion became slightly more representative of a demographic group's opinions. However, the group representativeness of these language models improved by a constant factor—meaning that even if there were representativeness improvements *within* a demographic group, there were still performance disparities *between* demographic groups. For instance, if a model originally aligned better with liberals than conservatives, steering makes it moderately more liberal and conservative, but it still represents liberals better.

Finally, on the consistency front, we found that models expressed a range of disparate opinions. This is reflective of reality since people often hold seemingly inconsistent or even contradictory beliefs. Still, it is interesting that language models reflect that reality.

> *...language models tuned with reinforcement learning from human feedback (...) are more aligned with left-leaning, liberal views.*

We still have a long way to go when it comes to developing mechanisms to evaluate language models. Our paper, while novel, also has limitations. It uses a multiple-choice question format to evaluate language models, while also assuming that users rely significantly on chatbots' answers to subjective questions and that the results of a language model poll can be compared to a poll of human beings. We also do not explore the important question of whether and when we even want language models to align with human opinions.

Policy Discussion

The focus on ChatGPT has made it more clear than ever that language models have major social, economic, and political ramifications. Policymakers have much to grapple with, ranging from the accessibility of these language models for different populations to the privacy and policy implications of companies gathering and training data for language models.

First and foremost, there is still much research to be done on the evaluation of language models. There are open questions about whether language models can replicate results from human experiments, such as in the social sciences, cognitive science, and economics. In addition, language model opinions that are prompted by providing multiple-choice options may not directly correspond to the behavior of the model when interacting with users in open-ended settings. Subjectivity in language model outputs is another key area for further work. How people judge the “correctness” of an answer given by a language model could depend on a variety of factors worth studying. Bias, fairness, toxicity, and other measurement variables are likewise important for public and policy reasons. There is great potential for policymakers to encourage or facilitate discussion and research on these issues.

> *The focus on ChatGPT has made it more clear than ever that language models have major social, economic, and political ramifications.*

Language model evaluation capabilities could provide policymakers, and potentially regulators, with the ability to quantitatively assess language model behavior and compare it to human preferences and opinions. But much more research progress is needed before language models can support policy analysis, algorithmic evaluation, and other activities to the level sought by some policymakers.

As language model development continues apace, evaluating these models is critical to understanding their impact on the public and mitigating potential harms.

The original article is accessible at Shibani Santurkar et al., "**Whose Opinions Do Language Models Reflect?**," *PMLR: Proceedings of the 40th International Conference on Machine Learning*, vol. 202 (July 2023): 29971-30004, https://proceedings.mlr.press/v202/santurkar23a.html.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Shibani Santurkar is a former postdoctoral scholar in computer science at Stanford University.

Esin Durmus is a former postdoctoral scholar in computer science at Stanford University.

Faisal Ladhak is a visiting student researcher in computer science at Stanford University.

Cinoo Lee is a PhD student in psychology at Stanford University.

Percy Liang is an associate professor of computer science and statistics and the director of the Center for Research on Foundation Models at Stanford University.

Tatsunori Hashimoto is an assistant professor of computer science at Stanford University.

HAI **Stanford University**
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**