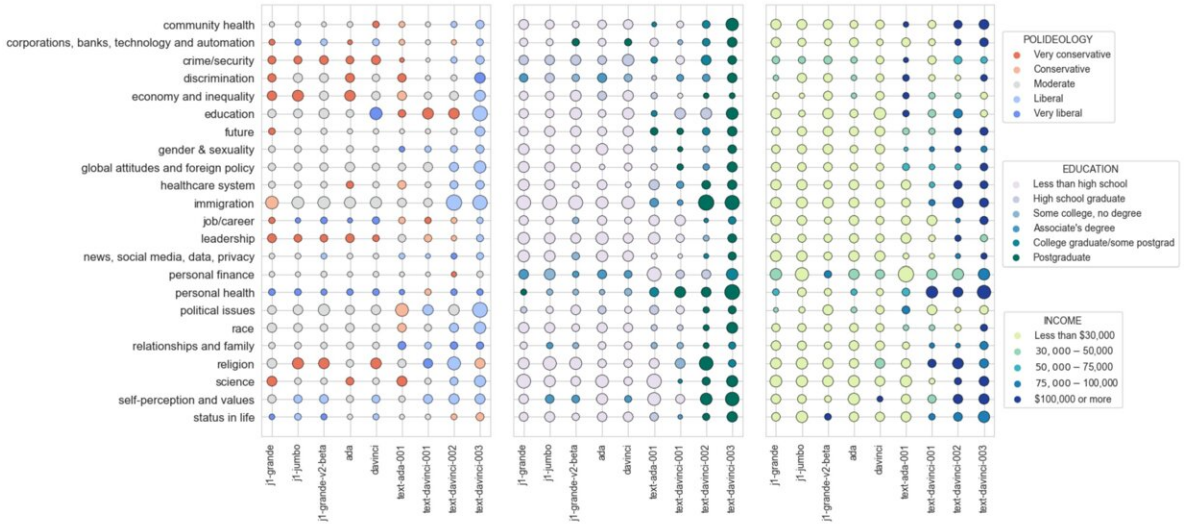
# Assessing political bias in language models

May 31 2023, by Andrew Myers




Consistency of different LMs (columns) across topics (rows) on different demographic attributes (panels). Each dot indicates an LM-topic pair, with the color indicating the group to which the model is best aligned, and the size of the dot indicates the strength of this alignment (computed as the ratio of the best and worst subgroup representativeness for that topic, see Appendix B.3 for details). We find significant topic-level inconsistencies, especially for base LMs, and strong educational attainment consistency for RLHF trained LMs. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2303.17548

The language models behind ChatGPT and other generative AI are trained on written words that have been culled from libraries, scraped from websites and social media, and pulled from news reports and speech transcripts from across the world. There are 250 billion such

words behind GPT-3.5, the model fueling ChatGPT, for instance, and GPT-4 is now here.

Now new research from Stanford University has quantified exactly how well (or, actually, how poorly) these models align with opinions of U.S. demographic groups, showing that language models have a decided bias on hot-button topics that may be out of step with general popular sentiment.

"Certain language models fail to capture the subtleties of human opinion and often simply express the dominant viewpoint of certain groups, while underrepresenting those of other demographic subgroups," says Shibani Santurkar, a former postdoctoral scholar at Stanford and first author of the study. "They should be more closely aligned."

In the paper, a research team including Stanford postdoctoral student Esin Durmus, Columbia Ph.D. student Faisal Ladhak, Stanford Ph.D. student Cinoo Lee, and Stanford computer science professors Percy Liang and Tatsunori Hashimoto introduces OpinionQA, a tool for evaluating bias in language models. OpinionQA compares the leanings of language models against public opinion polling.

As one might expect, language models that form sentences by predicting word sequences based on what others have written should automatically reflect popular opinion in the broadest sense. But, Santurkar says, there are two other explanations for the bias. Most newer models have been fine-tuned on human feedback data collected by companies that hire annotators to note which model completions are "good" or "bad." Annotators' opinions and even those of the companies themselves can percolate into the models.

For instance, the study shows how newer models have a greater-than-99 percent approval for President Joe Biden, even though public opinion

polls show a much more mixed picture. In their work, the researchers also found some populations are underrepresented in the data—those age 65 or older, Mormons, and widows and widowers, just to name a few. The authors assert that to improve credibility, language models should do a better job of reflecting the nuances, the complexities, and the narrow divisions of public opinion.

## Aligning to public opinion

The team turned to Pew Research's American Trends Panels (ATP), a benchmark survey of public opinion, to evaluate nine leading language models. The ATP has nearly 1,500 questions on a broad range of topics, stretching from science and politics to personal relationships. OpinionQA compares language model opinion distribution on each question with that of the general U.S. populace as well as the opinions of no fewer than 60 demographic subgroups, as charted by the ATP.

"These surveys are really helpful in that they are designed by experts who identify topics of public interest and carefully design questions to capture the nuances of a given topic," Santurkar says. "They also use multiple-choice questions, which avoid certain problems measuring opinion with open-ended questions."

From those comparisons, OpinionQA calculates three metrics of opinion alignment. First, representativeness assesses how aligned a language model is with the general population as well as against the 60 demographic cross sections ATP uses. Second, steerability tabulates how well the model can reflect the opinion of a given subgroup when prompted to do so. And third, consistency predicts how steady a model's opinions are across topics and across time.

## Wide variation

High-level findings? All models show wide variation in political and other leanings by income, age, education, etc. For the most part, Santurkar says, models trained on the internet alone tend to be biased toward less educated, lower income, or conservative points of view. Newer models, on the other hand, further refined through curated human feedback tend to be biased toward more liberal, higher educated, and higher income audiences.

"We're not saying whether either is good or bad here," Santurkar says. "But it is important to provide visibility to both developers and users that such biases exist."

Acknowledging that exactly matching the opinions of the general public could represent a problematic goal in itself, the developers of OpinionQA caution that their approach is a tool to help developers assess political biases in their models, not a benchmark of optimal outcomes.

"The OpinionQA dataset is not a benchmark that should be optimized. It is helpful in identifying and quantizing where and how language models are mis-aligned with human opinion and how models often don't adequately represent certain subgroups," Santurkar says. "More broadly, we hope it can spark a conversation in the field about the importance and the value of bringing language models into better alignment with public opinion."

The findings are published on the *arXiv* preprint server.

**More information:** Shibani Santurkar et al, Whose Opinions Do Language Models Reflect?, *arXiv* (2023). DOI: 10.48550/arxiv.2303.17548

Provided by Stanford University