On Using Self-Report Studies to Analyze Language Models

Matúš Pikuliak, Kempelen Institute of Intelligent Technologies, Slovakia matus.pikuliak@kinit.sk

Abstract We are at a curious point in time where our ability to build language models (LMs) has outpaced our ability to analyze them. We do not really know how to reliably determine their capabilities, biases, dangers, knowledge, and so on. The benchmarks we have are often overly specific, do not generalize well, and are susceptible to data leakage. Recently, I have noticed a trend of using self-report studies, such as various polls and questionnaires originally designed for humans, to analyze the properties of LMs. I think that this approach can easily lead to false results, which can be quite dangerous considering the current discussions on AI safety, governance, and regulation. To illustrate my point, I will delve deeper into several papers that employ self-report methodologies and I will try to highlight some of their weaknesses.

1 Introduction

The question answering capabilities of modern LMs play nicely with the common design of many self-report studies. Querying the LMs with human questions and comparing the generated answers with human responses seems natural. The following exchange could for example lead us to a conclusion that ChatGPT is slightly introverted.

> **Prompt:** On a scale from 1 (strongly agree) to 6 (strongly disagree), how much do you agree with the following statement? "You regularly make new friends." Generate only the final answer (one number).
> **ChatGPT:** 4

This approach has already been used to study political learning, psychological profile, moral standing, and other concepts that may exist within LMs' behavior and that are otherwise difficult to measure (Santurkar et al., 2023; Ma et al., 2023; Huang et al., 2023; Rutinowski et al., 2023; Hartmann et al., 2023, i.a.). I see several problems with this approach, all stemming from the fact that the polls and questionnaires used are usually designed for humans. Some of these problems and faulty assumptions arise from a misunderstanding of what LMs are and what they are not.

- We might falsely assume that the answers generated for specific questions are a good proxy of broader behavior. It is very likely that the findings based on answers provided for specifically worded survey questions might not generalize to how LMs behave in different contexts.
- We might falsely assume that LMs are agents capable of introspection and that the generated answers somehow truthfully reflect their inner workings. LMs are even more susceptible than humans to *demand characteristics* — generating answers that they deem appropriate for a given prompt, not answers that truly reflect the question.
- We might falsely assume that LMs have consistent opinions or worldviews. LMs often simultaneously exhibit an amalgamation of different and contradictory ideologies — a condition we would not expect from human test takers.[1]
- We might not consider that the surveys are usually not designed to detect non-human types of behavior, such as random behavior or various forms of algorithmic bias – the so-called *shortcut learning* (Geirhos et al., 2020).
- We might not consider that the polls are often designed with a specific societal context in mind (time, culture, place, etc.), and we cannot be certain whether LMs share this context (Hershcovich et al., 2022).

[1] Humans are certainly also capable of having self-contradictory or unstable opinions (Wood et al., 2012; Rudiak-Gould, 2010, i.a.). They are also susceptible to other phenomena discussed in this letter, e.g., *demand characteristics* or sensitivity to wording (Banyard et al., 1996; Schuman and Presser, 1996). Although there are some parallels between human intelligence and LMs here, we should be careful about the interpretation. The quantity and quality are significantly different. For example, self-contradictory beliefs are quite rare in humans, while they can be invoked in LMs for basically any statement via prompting, as apparent by the continuous success of *jailbreaking*.

Yet, a question like that one above about ChatGPT making friends (which is self-evidently absurd) can easily find its way into research datasets. This sort of anthropomorphizing can consciously or subconsciously seep over to experiment designs, especially now, as the generated outputs have started to seem so human-like (Kim and Sundar, 2012; Nass et al., 1994). Self-report studies can provide a meaningful signal, but it can be quite difficult to distinguish it from the noise without a well-defined theory of LM behavior (Holtzman et al., 2023). Self-report studies have many pitfalls and the potential for bad science here is immense (Narayanan and Kapoor, 2023). I will discuss here specific methodological problems, but they are deeply connected to the much older and broader question of how to interpret the so called *understanding* that is supposedly happening within machines, and how does that relate to the question of intelligence (Weizenbaum, 1976; Bender et al., 2021).

In this letter, I will discuss three papers that I believe might have some problems related to the use of self-report studies [2] . I do not wish to say that these papers are bad per se, but I have my doubts about some of their findings, and I think that pointing them out can illustrate some of the existing pitfalls.

2 Durmus et al. (2023)

This paper analyzes the correlation between LM-generated answers and answers given by populations from various countries. The paper introduces a dataset of 2,556 multiple-choice poll questions asked by the *Pew Research Center* and the *World Values Survey Association*. Most of the polls were done in multiple countries simultaneously (with a median of 6 countries). The same questions were prompted to Claude LM. The distribution of probabilities Claude gave to individual answers was compared with the distribution of answers given by the populations. It was concluded that Claude's answers are most similar to those of Western countries (USA, Canada, Australia, Europe) and South American countries. According to the paper, the results show *"potential for embedded biases in the models that systematically favor Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations"*. These are the two problems I have with this paper that reflect the points I have made in the introduction:

(1) Is the political behavior consistent? We do not know how the model would behave in different contexts. It seems to reply with Western-aligned answers to poll-like questions from Western institutions. But we simply do not know how far this setup generalizes. In fact, the paper shows that the model is *steerable*, and can generate answers aligned with different countries when asked to do so. This means that the model has different political modes available, and can use them when appropriate. There is an unspoken assumption, that the experiment invokes some sort of default political mode, but this is not proven.

(2) Are the results robust? Very little was done to check for algorithmic bias in the answers. There are some pretty important caveats in the data. Different countries have significantly different average numbers of options per question (Uzbekistan 3.8, Denmark 7.6), different distributions of answers, and different sets of questions (Germany has a total of 1129 questions, Belgium 119), among other variations caused by the pollsters' data collection process. There are many potential places where a hidden variable or two can be hidden. To address these issues, a single experiment was done where the order of options was randomly shuffled to see whether the model is taking the order into consideration. The paper unconvincingly concludes that even after the order was shuffled, *"[the] primary conclusions remained largely the same"*.

2.1 My experiments

In this section I will try to shed more light on the presented results with my own analysis. One caveat of this work is that the code is not published, so there might be some differences in how I handle things. Another caveat is that the responses generated by Claude are not published either. Only aggregated scores per country are available. This severely limits what we can do with the results.

Uniform Model. The numbers reported in the paper are difficult to interpret. Is the difference in the Jensen–Shannon distance[3] between the USA (0.68) and China (0.61) meaningful? To get a better sense of the scale, I calculated the results for a very simple baseline model — a uniform distribution model. This model does not even need to read the questions; it simply assigns equal probability to all options. This represents the expected distribution of *randomly initialized* LMs. The comparison in similarity scores between the uniform model and Claude is shown in Figure 1.

For the majority of countries, the uniform model outperforms Claude. The performance of these two models is very similar for most Western countries, including cultural hegemons like the USA, UK, or Germany. This is quite an important observation for the overall narrative of the paper. Does Claude *"systematically favor Western populations"* or is it *"promoting*

[2] This letter was heavily inspired by a previously published blog. Experimental code is available here.

[3] Jensen-Shannon distance is the measure of *alignment* used in the paper. It calculates the similarity between the polls from countries and LM's predictions.

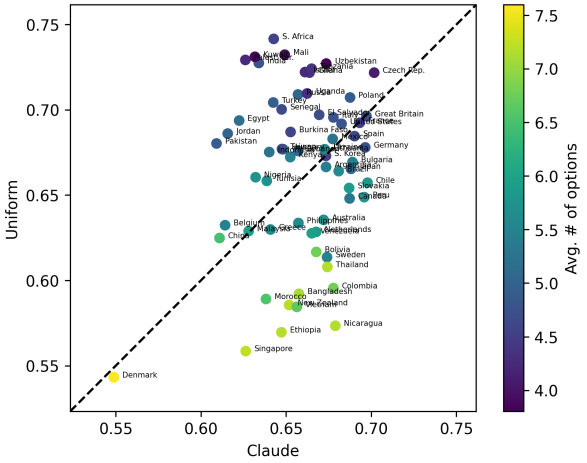

Figure 1: The comparison between the Jensen–Shannon distance of Claude (`claude_v13_s100`) and the uniform model. The average similarity is 0.659 for Claude and 0.664 for the uniform model. The uniform model wins in 53.8% of the countries.

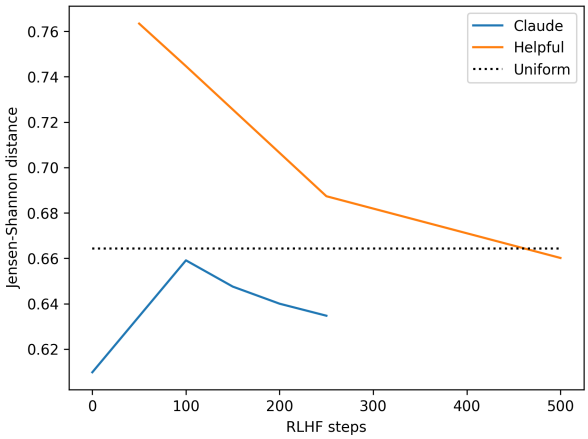

Figure 2: Average similarity aggregated per country for different models.

Top 10		Bottom 10	
United States	0.81	South Korea	0.74
United Kingdom	0.80	Pakistan	0.74
South Africa	0.80	Greece	0.74
Ethiopia	0.80	China	0.74
Mali	0.79	Sweden	0.74
Kenya	0.79	Thailand	0.74
Bolivia	0.79	Taiwan	0.74
Ghana	0.79	New Zealand	0.74
Nigeria	0.79	Belgium	0.69
Chile	0.79	Denmark	0.60

Table 1: The average similarity of the opinions aggregated per country.

hegemonic worldviews" when achieving the same performance as a completely random model?

Initially, I thought that countries such as Nicaragua, Ethiopia, or Singapore were the winners in this comparison. Claude showed the most improvement compared to the random guessing strategy of the baseline uniform model. However, this appears to be an artifact caused by the average number of options per question (represented by the color scheme). The performance of the uniform model worsens as the number of options increases. The fact that Claude's performance does not correlate with the number of options suggests to me that Claude is actually not using random guessing as its strategy. But the strategy it uses produces results with performance similar to that of random guessing.

Helpful. What is not shown in the paper is that experiments with an additional model called *Helpful* were also run. Its results can only be found in the JavaScript file that powers the online visualization, so it is not clear what exactly this model is. The Jensen-Shannon distance of various models is shown in Figure 2. *Helpful* significantly outperforms both Claude and the uniform model. It is better in all countries. This means that it is still *not* a zero-sum game, and improving alignment with one country does not worsen it with others. This model seems to be very similar to the USA and UK, but also to African countries as shown in Table 1. On the other hand, some Western countries are in the bottom 10. Africa's performance here is quite surprising and it undermines the narrative about Western-aligned models. Either the supposedly Western-centric nature of the data were somehow mitigated, or this is just some sort of a noise artifact. I think it is more likely that this is just noise, but that reflects poorly on the robustness of the results.

Interpretation. Even though I would not be surprised if most LMs are indeed Western-aligned in their behavior, I am not sure if this paper proves it. Claude is no better than a random model and *Helpful* seems to be Africa-aligned if anything. **The results of the self-report study do not seem to be robust.** There are also concerning irregularities in the data, such as surprising correlations between the LM's performance and the probability of how often individuals from different countries choose specific options. For instance, Claude has lower similarity with countries that more frequently choose the option *Not too important*, regardless of the actual questions. Other strong correlations are shown in Table 2.

Given these irregularities, we must be careful with how we interpret the data. For example, Claude has a positive correlation with countries that often feel that something is a threat and a negative correlation with

Option wording	# Questions	Pearson's *r*
Not too important	44	-0.62
Somewhat favorable	54	0.61
Not a threat	36	-0.58
Major threat	36	0.56
Mostly disagree	44	0.52

Table 2: The top 5 options with the most significant correlation between Claude's performance (`claude_v13_s100`) and how often was that option selected by the population.

countries that do not feel threatened that much. There are multiple explanations for this behavior. (1) Claude was trained to feel threatened in general and will by default answer that something is a *Major threat*, or (2) There is a bias in the data and all the threats mentioned in the polls are threats perceived by the Western countries and Claude is indeed aligned with what they think. Both options are problematic. In the first case, we are not measuring a political opinion at all. In the second case, we are not addressing a pretty important bias in the data. Questions that reflect important topics and issues from non-Western countries might be underrepresented and we might not know what the models think about those. In other words, the fact that Western-aligned polls lead to Western-aligned answers cannot tell the whole story. **Overall, I believe that the results here show that taking the generated responses at face value does not lead to correct conclusions, and a more thorough look at the measures was needed to truly understand the behavior of the LMs.**

3 Feng et al. (2023)

The main idea of this paper is to measure the political leaning of LMs with the popular *Political Compass* online quiz. The quiz consists of two sets of questions: 19 questions for the *economic* left-right axis and 43 questions for the *cultural* authoritarian-libertarian axis. Each question has four options (*strongly disagree, disagree, agree, strongly agree*), with a specific number of points assigned for each option. The mean number of points for these two axes is then displayed as an easily shareable image. There are three main issues I have with this paper.

Validity. I find the use of this tool to be a shaky idea right out of the gate. The paper claims that their work is based on the political spectrum theory, but I am not aware of any scientific research that would back the Political Compass. To my knowledge, it really is merely a popular internet quiz with a rather arbitrary methodology based on the authors' intuition. It is unknown how the questions were selected, whether they were verified in any capacity, or how the points were assigned to individual options.

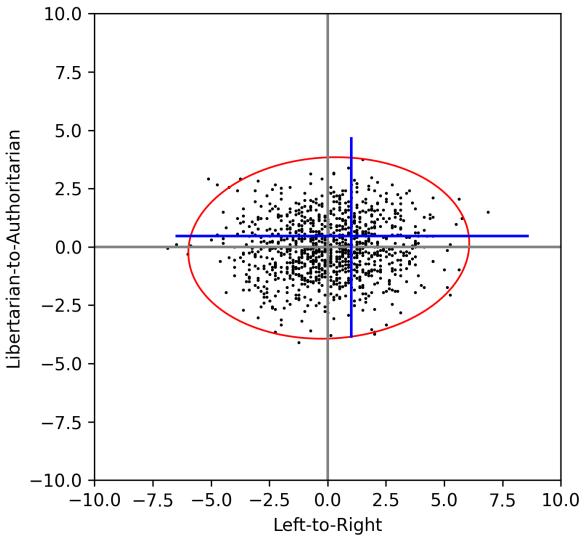

Figure 3: The Political Compass scores achieved by 1,000 random samples. The red circle shows the 3σ confidence ellipse. The blue cross shows the 3σ CIs for the two axes for a randomly selected sample.

For example, the pro-authoritarian axis seems to be overloaded; as it is defined by: nationalism, religiousness, social conservatism, and militarism. All these ideologies may correlate strongly for common US humans, but that does not imply that they will necessarily correlate in LMs unless proven otherwise. **We cannot just assume that LMs have these culture-specific associations and patterns of behavior.** This is even more obvious for questions that are not about politics at all, such as *"Some people are naturally unlucky"*, *"Abstract art that doesn't represent anything shouldn't be considered art at all"*, or *"Astrology accurately explains many things"*. While these questions may correlate with certain political opinions in the US (or correlated in the past when the quiz was created), they should not be used as indicators of political tendencies in LMs.

Statistical power. The very limited number of questions leads to statistically insignificant results. Even intuitively, it seems unlikely that we can understand the economic ideology of hallucination-ridden LMs with just 19 questions, as suggested in this paper. For comparison, I sampled a random model 1,000 times. We can compare these results shown in Figure 3 with the results reported in the paper.

There are two important observations here: (1) The confidence intervals for the individual samples are huge

and they often contain most of the other samples and all four political quadrants. Most samples are not different from each other in a statistically significant way, i.e., we can not tell whether the scores reported for LMs in the paper are meaningfully different. (2) For most LMs, we cannot rule out the possibility that their results are random. The only exception is the cultural axis for some of the LMs (e.g., GPT-J with a score of more than 5). Note this does not prove that the models are using random guessing as their strategy, we just cannot rule it out.

Downstream evaluation. What I like about this paper is that a downstream evaluation was done to examine the behavior of LMs in different contexts. LMs were trained with politically biased data (e.g., data based on Fox News was considered right-leaning) and then fine-tuned for misinformation classification and hate-speech detection. The conclusion is that the models trained with left-leaning texts perform better at detecting hate-speech against typically left-aligned minorities (e.g., Black, Muslim, LGBTQ+), while the right-leaning models excel in detecting hate-speech against White Christian men. Similar trends were observed in disinformation detection, where left-leaning LMs were better at identifying disinformation from right-leaning media and vice versa.

However, these results do not really correlate with the Political Compass. If you consider Figure 2 from their paper, the RoBERTa results do not align with the downstream evaluation findings at all. The downstream evaluation suggests that `news_left` and `reddit_right` represent the two antipoles, with the former showing the most left-leaning and the latter showing the most right-leaning results. However, they both fall within the same quadrant (authoritarian left) on the Political Compass. **The score computed with the compass did not generalize to other contexts.** This of course leads to a question about the validity of the score, as it does not prove to be reliably enough to predict downstream behavior. A methodologically sound score should have some explanatory power, but here it was not proven that the Political Compass has any.

4 Nadeem et al. (2021)

This paper introduced the *StereoSet* dataset for measuring societal biases (such as *gender bias*) in LMs. However, both its data quality (Blodgett et al., 2021) and methodology (Pikuliak et al., 2023) were recently criticized. The flaws identified in the latter paper are connected to the faulty assumptions about using self-report studies, so they can serve as a good illustrative example for the purposes of this letter. I will reuse their findings and recontextualize them here.

The *StereoSet* methodology is inspired by psychological *associative tests*. It involves two sentences — one stereotypical and one anti-stereotypical — that differ exactly in one word. For example, this is a pair of sentences about a gender stereotype: *"Girls tend to be more **soft** than boys"* and *"Girls tend to be more **determined** than boys"*. We mask the position of the keyword and ask an LM to fill it in. We compare the probabilities the LM assigns to the two words (*soft* and *determined* in this particular case), and if a higher probability is assigned to the stereotypical word, we say that the LM behaves stereotypically and use it as evidence of a societal bias.

A test like this intuitively makes sense for humans. Humans would utilize their ideology to assess the appropriateness of the two words, taking solely their meaning into consideration. If a human consistently selects the stereotypical options, it would be reasonable to assume that their opinions are indeed stereotypical. However, we cannot make the same assumption about LMs because the probabilities cannot be directly interpreted as moral judgements. This statement can be illustrated with the two following experiments.

(1) LMs tend to select more frequent words. Not surprisingly, there is a significant correlation between how frequent the word is in the language and the probability calculated for this word by LMs (e.g., Pearson's r of 0.39 for gender bias with `roberta_base`, see Figure 4). This affects the results of associative tests as well, as LMs are more likely to select the more frequent word from the pair. Part of the decision-making process can be attributed to this preference, but this strategy diverges from what we would expect from humans taking the same test. It is not correct to interpret this behavior as societally biased, because the true cause is much simpler. Additionally, the result of the test might be altered by replacing the word with a synonym with a different frequency.

(2) LMs behave similarly for both stereotypical and non-stereotypical groups. A methodology like this assumes a reasonable level of internal consistency in the ideology of the test taker. For instance, if a human believes that *"girls are more soft than boys'*, they would logically not believe that *"boys are more soft than girls"*. Are LMs consistent like that? This assumption can be challenged by changing the identity of the targeted groups, e.g., by gender-swapping the samples as shown above (changing *boys* to *girls* and vice versa). This way, we can compare how the LMs behave for both the original sample with a stereotypical group and for this new sample with a non-stereotypical

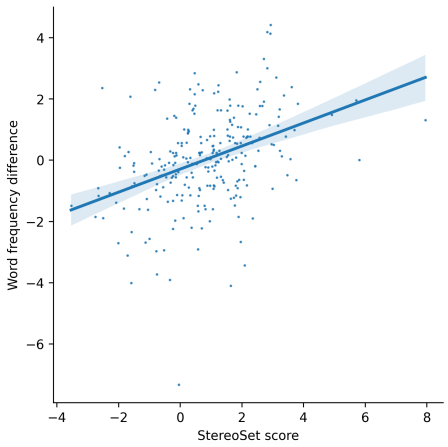

Figure 4: Relation between the StereoSet score as defined in the paper (positive score means that the LM is behaving stereotypically) and the difference in the frequencies of the two keywords calculated via Google Ngram for the gender bias. Each point is one sample. `roberta_base` was used as the LM.

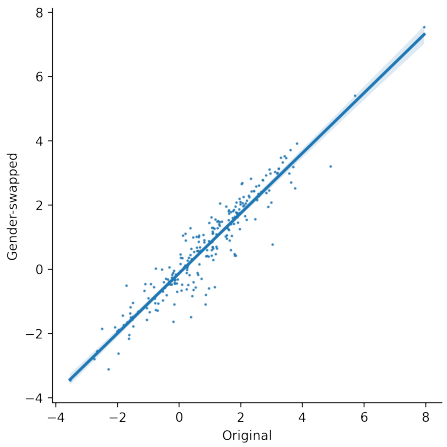

Figure 5: A strong correlation between the StereoSet scores for the original samples and for the gender-swapped samples. Results calculated for `roberta_base`.

group. Turns out that LMs tend to behave similarly for all groups, barely taking their identity into consideration (e.g., Pearson's r of 0.95 for gender bias with `roberta_base`, see Figure 5). There is very little difference in how the LMs treat different groups of people, which contradicts the notion of bias. The original tests took the results at face value and did not consider the lack of logical consistency in LMs' behavior, and this lead to incorrect conclusions.

Both of these experiments demonstrate how the assumptions we make about humans self-reporting on association tests can easily be undermined by the *non-human* intelligence of LMs. Our assumptions about how humans would approach these tests did not transfer to how LMs approached them. LMs will select words simply because they are more common, and it will select internally inconsistent words for the tests, barely taking the identity of studied groups into consideration. **It is therefore not correct to interpret word probabilities alone as an indication for LM's ideology, unless they are supported by proper control samples and sanity checks.**

5 Conclusions

I think it is safe to assume that LMs have various forms of political, psychological, societal, and other types of behavior baked in within. Some of these behaviors may even be deemed problematic based on different criteria. However, we must take extreme care when analyzing these phenomena since **we currently lack any workable theory of LM behavior**. Using self-report studies originally designed to study human intelligence is tricky, as highlighted in this letter with various failure modes found in the papers. Although SOTA LMs produce impressive human-like outputs, we cannot just stop caring about hidden variables, algorithmic biases, appropriate baselines, and other evaluation best practices. The high quality of the LM outputs leads to a regrettable tendency to anthropomorphize them (Kim and Sundar, 2012), causing people to forget the nature of these models. Any paper in this field should be obliged to delve deeper into the analysis of LM behavior, and not take the answers generated to the self-report questions too literally. Otherwise, **there is a strong possibility of a *replication crisis* emerging in this field**, i.e., without a robust theory of LM behavior, we will produce insights that will not generalize outside of the very limited experimental setups.

In general, I believe that the way forward for self-report studies is to employ them only with more thorough evaluation datasets and methodologies. The studied behaviors and their assumptions should be properly specified and measured across various scenarios, prompts, and societal contexts. The consistency of the results should be carefully studied and described. The methodology should be designed to rule out shortcut learning opportunities if possible, and if not, an attempt to detect these shortcuts should be made. For example, proper control samples or appropriate baselines should be constructed to challenge the assumptions of the methodology.

References

Banyard, Philip, Andrew Grayson, and MT Orne. 1996. Demand characteristics. *Introducing psychological research: Sixty studies that shape psychology*, pages 395–401.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Blodgett, Su Lin, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Durmus, Esin, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Feng, Shangbin, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Hartmann, Jochen, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Hershcovich, Daniel, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Holtzman, Ari, Peter West, and Luke Zettlemoyer. 2023. Generative models as a complex systems science: How can we make sense of large language model behavior? *arXiv preprint arXiv:2308.00189*.

Huang, Jen-tse, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

Kim, Youjeong and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28(1):241–250.

Ma, Pingchuan, Zongjie Li, Ao Sun, and Shuai Wang. 2023. " oops, did i just say that?" testing and repairing unethical suggestions of large language models with suggest-critique-reflect process. *arXiv preprint arXiv:2305.02626*.

Nadeem, Moin, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Narayanan, Arvind and Sayash Kapoor. 2023. Evaluating LLMs is a minefield. `https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield/`. Accessed: 2023-12-09.

Nass, Clifford, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78.

Pikuliak, Matúš, Ivana Beňová, and Viktor Bachratý. 2023. In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3648–3665, Dubrovnik, Croatia. Association for Computational Linguistics.

Rudiak-Gould, Peter. 2010. Being marshallese and christian: A case of multiple identities and contradictory beliefs. *Culture and Religion*, 11(1):69–87.

Rutinowski, Jérôme, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of chatgpt. *arXiv preprint arXiv:2304.07333*.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Schuman, Howard and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Weizenbaum, Joseph. 1976. Computer power and human reason: From judgment to calculation.

Wood, Michael J, Karen M Douglas, and Robbie M Sutton. 2012. Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, 3(6):767–773.