Questioning the Survey Responses of Large Language Models

Ricardo Dominguez-Olmedo[1,2] **Moritz Hardt**[1,2] **Celestine Mendler-Dünner**[1,2,3]

[1]Max-Planck Institute for Intelligent Systems, Tübingen
[2]Tübingen AI Center
[3]ELLIS Institute Tübingen
{rdo,hardt,cmendler}@tuebingen.mpg.de

Abstract

Surveys have recently gained popularity as a tool to study large language models. By comparing survey responses of models to those of human reference populations, researchers aim to infer the demographics, political opinions, or values best represented by current language models. In this work, we critically examine this methodology on the basis of the well-established American Community Survey by the U.S. Census Bureau. Evaluating 43 different language models using de-facto standard prompting methodologies, we establish two dominant patterns. First, models' responses are governed by ordering and labeling biases, for example, towards survey responses labeled with the letter 'A'. Second, when adjusting for these systematic biases through randomized answer ordering, models across the board trend towards uniformly random survey responses, irrespective of model size or pre-training data. As a result, in contrast to conjectures from prior work, survey-derived alignment measures often permit a simple explanation: models consistently appear to better represent subgroups whose aggregate statistics are closest to uniform for any survey under consideration.

1 Introduction

Surveys have a long tradition in social science research as a means for gathering statistical information about the characteristics, values, and opinions of human populations [Groves et al., 2009]. Insights from surveys inform policy interventions, business decisions, and science across various domains. Surveys typically consist of a series of well-curated questions in a multiple-choice format, with unambiguous framing and a set of answer choices carefully selected by domain experts. Surveys are then presented to groups of individuals and their answers are aggregated to gain statistical insights about the populations that these groups of individuals represent.

Many established survey questionnaires together with the carefully collected answer statistics are publicly available. Machine learning researchers have identified the potential benefits of building on this valuable data resource to study large language models (LLMs). Survey questions offer a way to systematically prompt LLMs, and the aggregate statistics over answers collected by surveying human populations serve as a reference point for evaluation. As a result, the use of surveys has recently gained popularity for studying LLMs' biases [Santurkar et al., 2023, Durmus et al., 2023]. Also prompting LLMs with survey questions, researchers in the social sciences have explored using LLMs to emulate the survey responses of human populations [Argyle et al., 2023, Lee et al., 2023]. If effective proxies, simulated responses could augment or replace the expensive data collection process involving human subjects and provide insights into subpopulations that are otherwise hard to reach.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

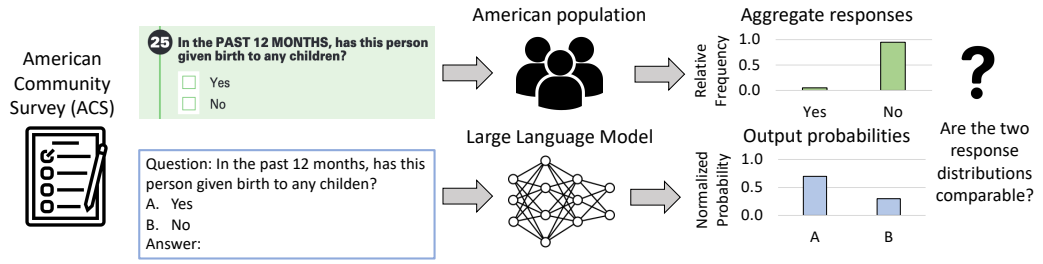


Figure 1: We prompt language models with questions from the American Community Survey (ACS). We systematically compare models' survey responses to those of the U.S. Census.

It is tempting to prompt LLMs with survey questions, due to their syntactic similarity to question answering tasks [Brown et al., 2020, Liang et al., 2022]. However, it is a priori unclear how to interpret their answers. Rather than knowledge testing, surveys seek to elicit aggregate statistics over individuals, providing an unbiased view on the properties of the population they are targeting. The quality of survey data hinges on the validity and robustness of the conclusions that can be drawn from it. Clearly, running a survey on LLMs is different from interrogating humans and thus it comes with distinct challenges. While much research has gone into carefully designing surveys to ensure faithful human responses, it is unclear whether prompting LLMs with the same surveys satisfies similar premises out-of-the-box. We devote this work to gain systematic insights into the survey responses of LLMs, what we can expect to learn from them, and to what extent they resemble those of human populations.

1.1 Our work

The basis of our investigation is the American Community Survey[1] (ACS), a demographic survey conducted by the U.S. Census Bureau at a national level, on a yearly basis. We curate a questionnaire composing of 25 multiple choice questions from the 2019 ACS. We prompt 43 language models of varying size with these questions, individually and in sequence, and we record their probability distribution over answers. Based on the collected data, we investigate the following two questions: *What can we infer about LLMs, and the data they have been trained on, from their survey responses? Does the data generated by prompting models to answer the ACS questionnaire qualitatively resemble the census data collected by surveying the U.S. population?* See Figure 1.

We start by inspecting models' distributions over answers to individual survey questions when the questions are asked independently. We observe that the entropy of response distributions differs substantially across models of varying size. Entropy tends to increase log linearly with model size, and it is preserved across different questions asked. We find that this differences arise because strong ordering and labeling biases confound models' answers. In fact, after adjusting for such systematic biases through randomized choice ordering, we find that response distributions are very similar across models and tend to correspond to highly balanced answers.

Comparing models' adjusted responses to those of the U.S. census population, we find that natural variations in entropy across questions are not reflected in the responses. Instead, on average across questions, models' responses are no closer to the census population, or the population of any state within the US, than to a fixed uniform baseline. This qualitative difference between model responses and human data puts into question the insights that can be gained from such comparisons. We find that even after instruction-tuning this trend persists, and model responses have consistently higher entropy than any human population we compare to, independent of the survey used. Only for models of size larger than 70 billion parameters we can recognize a trend that the divergence between model responses and the census data decreases after instruction-tuning.

With these insights in mind, we inspect conjectures from prior work related to survey derived alignment metrics, that is, that differences in similarity between models' and populations' responses might be attributable to certain demographics being better represented in the training data. Instead, our results suggest a much simpler explanation: the relative alignment of model responses with different

[1] https://www.census.gov/programs-surveys/acs

demographic subgroups can be explained by the entropy of the subgroups' responses, irrespective of the data or training procedure employed to train the model. We demonstrate this beyond the ACS on other surveys considered by prior work. As such, our findings provide important context to prior studies that employ surveys to examine the biases of LLMs.

More broadly, our findings suggest caution when treating language models' survey responses as a faithful representation of any human population, at least a present time, as it could lead to potentially misguided conclusions about alignment.

1.2 Related work

Despite the syntactical similarities, there are important differences between evaluating LLMs on the basis of their survey responses and traditional question answering evaluations [Liang et al., 2022]. Question answering (QA) tasks predominantly serve the purpose of knowledge testing [e.g., Kwiatkowski et al., 2019, Rajpurkar et al., 2016, Talmor et al., 2019, Mihaylov et al., 2018]. In such setting, a language model's answer to some unambiguous input question is extracted by computing its most likely completion. Similarly, for questions that lack a clear answer (e.g., "Angela and Patrick are sitting together. Who is an entrepreneur?") models' most likely response have been used to investigate various biases of LLMs [Li et al., 2020, Mao et al., 2021, Perez et al., 2023, Abid et al., 2021, Jiang et al., 2022].

When evaluating LLMs on the basis of survey questions, the focus is not on the model's most likely completion but rather on the probability distribution that the model assigns to various answer choices. For example, not whether the model is more likely to answer "Yes" than "No" to a given survey question, but the normalized probability assigned to each of the two answer choices. See Figure 1. More concretely, Santurkar et al. [2023] study LLMs' answer distributions for multiple-choice opinion polling questions, measuring their similarity to those of various U.S. demographic groups. They extract models' answer distributions from the next token probabilities corresponding to each answer choice. Subsequent works employ a similar methodology but instead consider transnational opinion surveys [Durmus et al., 2023, AlKhamissi et al., 2024] and moral beliefs surveys [Scherrer et al., 2024]. We adopt this popular methodology to systematically investigate the properties of models' answer distributions on the basis of a well-established demographic survey.

Instead of asking questions individually, Hartmann et al. [2023], Rutinowski et al. [2023], Motoki et al. [2023], Feng et al. [2023] sequentially prompt language models to answer entire political compass or voting advice questionnaires. Rather than aggregating answers into a political affinity score, our focus is instead on examining whether models' responses qualitatively resemble those of human populations. We discuss this sequential generation setting in detail in Appendix F.

Lastly, there is an emerging body of research that integrates LLMs into computational social science [Ziems et al., 2024]. This includes tasks such as taxonomic labeling, where language models are employed for tasks such as opinion prediction [Kim and Lee, 2023, Mellon et al., 2022], and free-form coding, where language models are used to generate explanations for social science constructs [Nelson et al., 2021]. Recent studies have also investigated the feasibility of using LLMs to simulate human participants in psychological, psycholinguistic, and social psychology experiments [Dillion et al., 2023, Aher et al., 2023], or as proxies for specific human populations in social science research [Argyle et al., 2023, Lee et al., 2023, Sanders et al., 2023] and economics [Brand et al., 2023, Horton, 2023]. Within this context, our work suggests caution in relying on the survey responses of LLMs to elicit synthetic responses that resemble those of human populations and highlights potential pitfalls.

2 Surveying language models

We employ the de-facto standard methodology to survey language models introduced by Santurkar et al. [2023]. For every survey question, we generate a prompt containing the multiple-choice question and we collect language models' probability distribution over answer choices. Formally, for a given model m and survey question q we define the model's *survey response* as a categorical random variable R_q^m which can take on k_q values corresponding to the number of answer choices to question q. The respective answer distributions are then contrasted with those of human populations align various dimensions. The overall setup is illustrated in Figure 1.

Prompting. We determine the event probabilities of R_q^m by prompting model m as follows:

1. We construct an input prompt of the form "`Question: <question> \n A. <choice 1 >\n B. <choice 2> \n ... <choice` k_q`> \n Answer:`".
2. We query language models with the input prompt and obtain their output distribution over next-token probabilities. We select the k_q output probabilities corresponding to each answer choice (e.g., the tokens "A", "B", etc.), and we renormalize to obtain the probability distribution over survey answers. [2].

The chosen style of prompt is standard for question answering tasks [Hendrycks et al., 2021], used in OpinionQA [Santurkar et al., 2023], and follows the best practices for social science research recommended by Ziems et al. [2024]. For completeness we perform several prompt ablations, including the prompt variations used by Argyle et al. [2023], Santurkar et al. [2023] and Durmus et al. [2023]. We find our take-aways to be robust to such changes, see Appendix D. However, note that our goal is not to engineer better prompts, but to critically examine popular scientific practices.

Survey questions. We use a representative subset of 25 multiple-choice questions from the 2019 ACS questionnaire. We denote the set of questions by Q. The questions cover basic demographic information, education attainment, healthcare coverage, disability status, family status, veteran status, employment status, and income. We generally consider the questions and answers as they appear in the ACS questionnaire. Figure 1 depicts an example question. We refer to Appendix A.1 for our list of questions and the exact framing we used for each question.

Models surveyed. We survey 43 language models of size varying from 110M to 175B parameters: the base models GPT-2 [Radford et al., 2019], GPT-Neo [Black et al., 2021], Pythia [Biderman et al., 2023], MPT [MosaicML, 2023], Llama 2 [Touvron et al., 2023],Llama 3 [Dubey et al., 2024] and GPT-3 [Brown et al., 2020]; as well as the instruct variants of MPT 7B and GPT NeoX 20B, the Dolly fine-tune of Pythia 12B [Databricks, 2023], Llama 2 Chat, Llama 3 Instruct, the text-davinci variants of GPT-3 [Ouyang et al., 2022], and GPT-4 [OpenAI, 2023].

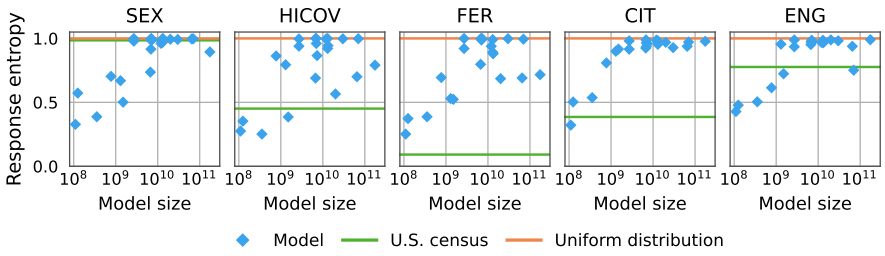
Reference data & evaluation. We use the responses collected by the U.S. Census Bureau when surveying the U.S. population as our reference data. In particular, we use the 2019 ACS public use microdata sample[3] (henceforth census data). The data contains the anonymized responses of around 3.2 million individuals in the United States. For each survey question $q \in Q$, we denote the census' population-level response as a categorical random variable C_q whose event probabilities are the relative frequency of each answer choice among survey respondents. We use U_q to denote the uniform distribution over answers. Given these two reference points, we evaluate language models' responses R_q^m along two dimensions:

- We use *entropy* to measure the degree of variation in models' responses. We denote the entropy of a random variable R as $H(R)$. To meaningfully compare the entropy of responses across questions with varying number of choices k_q, we report normalized entropy, that is, the entropy relative to the uniform distribution. $H(R_q^m) = 1$ implies that model m's survey response to question q is uniformly distributed (i.e., $H(U_q) = 1$).
- We use the *Kullback–Leibler (KL) divergence* to measure the "similarity" between two distributions over answers. We write $\mathrm{KL}(R_q^m \parallel C_q)$ for the KL divergence between the response distribution R_q^m of model m to question q and the corresponding aggregate response distribution C_q observed in the census data. The larger the KL distance between two distributions, the more dissimilar the two distributions are.

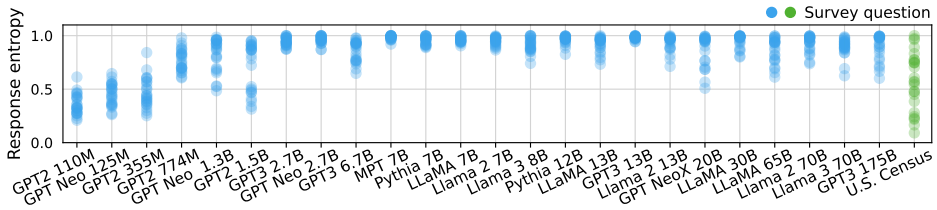
Note that the KL divergence between any distribution and the uniform distribution corresponds to the entropy difference. For normalized entropy this yields $\mathrm{KL}(C_q \parallel U_q) = k_q(1 - H(C_q))$.

[2]For OpenAI's models, we only have access to the top-5 next-token log probabilities through the OpenAI API. In this case, we assign to the unseen probabilities (if any) the minimum between the remaining probability mass and the smallest observed probability, following the methodology of Santurkar et al. [2023]

[3]https://www.census.gov/programs-surveys/acs/microdata



(a) Entropy of base models' responses, for five of the ACS questions.



(b) Entropy of base models' responses to the ACS, ordered by model size.

Figure 2: Entropy of model responses across the ACS questions for naive prompting. Entropy of models' responses (◆) tends to increase log-linearly with model size, irrespective of the underlying response entropy observed in the U.S. census (–).

Randomized choice ordering. For several investigations we survey models under randomized choice ordering. This means, for a given question q, we prompt models with different permutations of the answer choice ordering, i.e., the assignment of answers (e.g., "male", "female") to choice labels ("A", "B", etc), while the choice labels are kept in alphabetic order. We evaluate models' survey responses under all possible choice orderings and we use $\bar{R}_q^m$ to denote the expected distribution over answers and $\bar{O}_q^m$ to denote the expected distribution over selected choice labels. For questions with more than 6 answers we evaluate a maximum of 5000 permutations. For OpenAI's models we evaluate up to 50 permutations due to the costs of querying the OpenAI API. This distinction serves to decouple a model's tendency towards picking a particular answer from its tendency towards picking a particular choice label. In the following we refer to the expected survey response $\bar{R}_q^m$ under uniformly distributed choice ordering as the *adjusted* survey response. We will come back to this in Section 4.

3 Systematic biases in models' survey responses

We start by surveying the base pre-trained models. We present survey questions independently of one another, showing the answer choices in the same order as the ACS.

For a first investigation, we consider the normalized entropy of models' responses to the "SEX", "HICOV", and "FER" questions. The SEX question inquiries about the person's sex, encoded as male female, the HICOV question inquiries whether the person is currently covered by any health insurance plan, and the FER question inquires whether the person has given birth in the past 12 months. When surveying the U.S. population, these three questions elicit responses with very different entropy; responses to the SEX question are almost uniformly distributed, whereas most people answer "No" to the FER question. In contrast, as shown in Figure 2(a), the entropy of models' responses to these three questions are surprisingly similar. In particular, we find that the entropy of models' responses tends to increase log-linearly with model size, independent of the question asked. This trend is consistent across all ACS survey questions, see Figure 8 in Appendix B.1.

For a broader picture, we illustrate models' response entropy across all survey questions in Figure 2(b). The blue dots represent models' responses to individual questions, and the green dots represent the entropy of the responses of the U.S. census. We order models by size. We observe that the entropy of responses of the U.S. census greatly varies across questions. In contrast, for any given model, the entropy of its responses varies substantially less so.

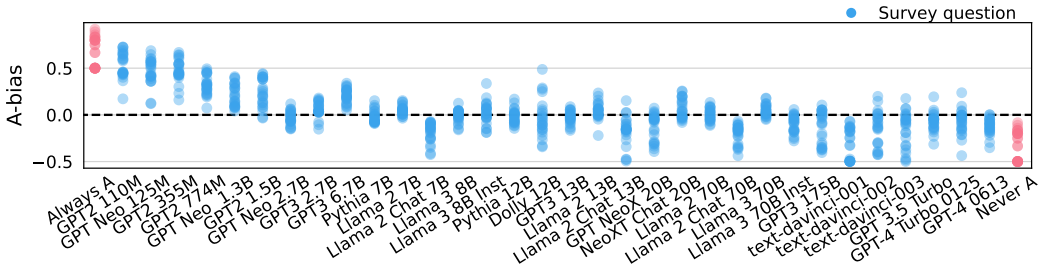


Figure 3: A-bias of in model responses across ACS questions. Each dot corresponds to one of the 25 questions. Models are ordered by size. As a reference, the extreme points illustrate A-bias for a model that always answers 'A' and a model that never answers 'A'. All models suffer from substantial A-bias.

Overall, we find that models' response distributions seem to be widely independent of the survey question asked, and variations across models are much larger than variations across questions. This lead us to suspect that observed differences across models might arise mostly due to systematic biases.

3.1 Testing for systematic biases: A-bias

It is well-known that language models' most likely answer to multiple-choice questions can change depending on seemingly minor factors such as the ordering of few-shot examples [Zhao et al., 2021, Lu et al., 2022] or the ordering of answer choices [Robinson and Wingate, 2023a]. We are interested in the extent to which changes in choice ordering affect a model's output *distribution over answers*.

We start by measuring *A-bias*: the tendency of a model towards picking the answer choice labeled "A". In particular, we seek to study the extent to which the strength of this bias explains the differences in responses observed across models. For an unbiased model that outputs the same answer distribution irrespective of choice ordering, the expected choice distribution $\bar{O}_q^m$ under randomized choice ordering would match precisely the uniform distribution (e.g., P("A") = P("B") = 0.5). We define a model's A-bias as its absolute deviation from this unbiased baseline:

$$\text{Abias}_q^m := \text{P}(\bar{O}_q^m = \text{“}A\text{”}) - 1/k_q \tag{1}$$

We measure A-bias for each question q and model m. Results are illustrated in Figure 3. We again sort models by their size. We observe all models exhibit substantial A-bias. However, models in the order of a few billion parameters or fewer consistently exhibit particularly strong A-bias, and tend towards mono answers. We additionally observe that the strength of A-bias in instruction or RLHF tuned models is similar to that of base models, see Appendix B.2. A plausible explanation for small models exhibiting strong A-bias is that the ability to answer MMLU-style multiple-choice questions only emerges for models of sufficient scale [Dominguez-Olmedo et al., 2024].

We investigate other types of labeling and position bias (e.g., last-choice bias) in Appendix C. Overall, we find a strong tendency of LLMs to pick up on spurious signals in the way that answers are ordered and labeled, rather than their semantic meaning. Notably, in contrast to the primacy bias observed in humans [Groves et al., 2009], we find that models exhibit substantial A-bias even when randomizing the position of the "A" choice. Our findings are consistent with the concurrent work of Tjuatja et al. [2023], which similarly finds that models' response biases to multiple-choice survey questions are generally not human-like. The orthogonal work of Wang et al. [2024] additionally shows that models' responses to multiple-choice survey questions may not consistently reflect their free-form outputs.

In summary, we find that systematic biases confound models' answer distributions. This makes it challenging to draw robust conclusions about inherent properties of LLMs, such as the opinions or populations they best represent. For example, simply reversing the order of answers to the "SEX" question could lead to GPT-2 seemingly representing a population where females are significantly over-represented, whereas a reverse conclusion would be drawn when using the standard answer order. While much research went into designing the ACS to elicit faithful answers and eliminate systematic biases when surveying human populations, simply using the same question framing does not protect against the systematic response biases that language models exhibit.

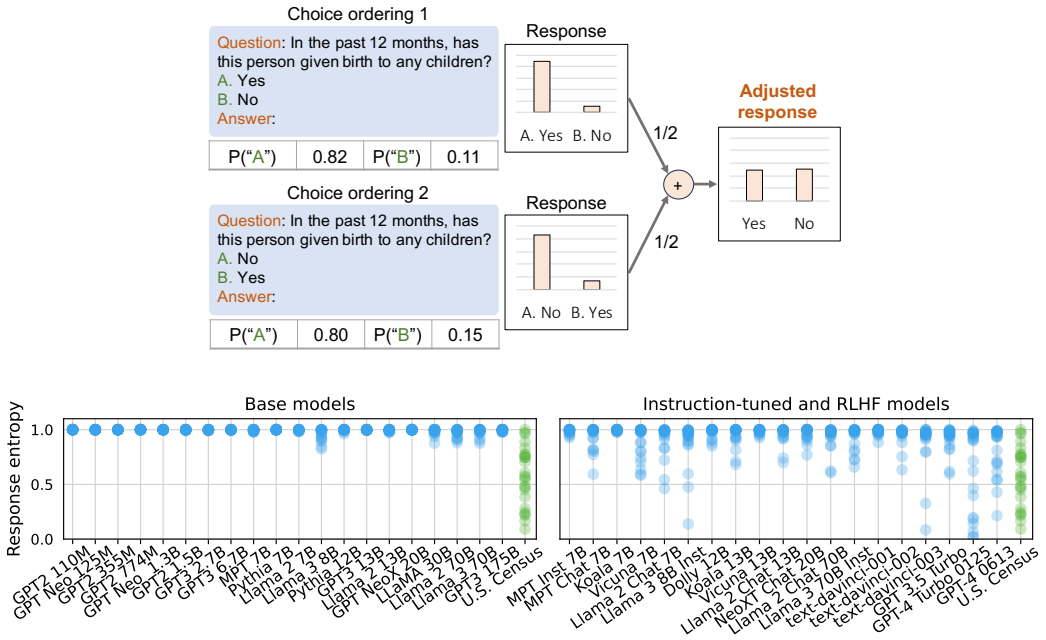


Figure 4: Entropy of model responses after adjustment. *(top)* Illustration of how adjustment is performed. We average models' responses over all possible answer orderings. *(bottom)* Entropy of models' responses after adjustment. Entropy of base models' responses is close to 1 (i.e., uniform). Instruction tuned-models exhibit substantially higher variations in entropy across questions.

4 Inspecting adjusted responses

To eliminate confounding due to labeling and ordering biases, we survey models under randomized choice ordering, borrowing an established methodology to adjust for ordering biases of all kinds in survey research [Groves et al., 2009]. Also a recent work in LLM research adopts this methodology [Robinson and Wingate, 2023b]. In the following, we refer to the expected response after answer choice randomization as the *adjusted* response.

In Figure 4 we plot the normalized entropy of models' adjusted responses for the ACS questions considered. First focusing on base models, and comparing the results to Figure 2(b) we find that after adjustment, 1) the variations in responses' entropy across survey questions are very small, 2) we no longer observe the trend of the entropy of model responses increasing log-linearly with model size. In fact, models' survey responses have a normalized entropy of approximately 1 irrespective of model size or survey question asked. This validates our initial hypothesis that, without adjustment, variations in responses across base models arise predominantly due to systematic biases such as A-bias, rather than the content of the survey questions asked.

4.1 Effect of instruction tuning

We now evaluate language models that have been fine-tuned with instructions and/or human preferences, henceforth "instruction-tuned models". In the right plot of Figure 4 we show the normalized entropy of instruction-tuned models' ACS survey responses after adjustment. We observe that instruction tuned-models all exhibit substantially higher variations in entropy across questions compared to base models. But in general, the entropy of their responses remains higher than the entropy of the census responses. Interestingly, as we will see, although deviating more from uniform, model responses do not tend to be closer to the U.S. census responses.

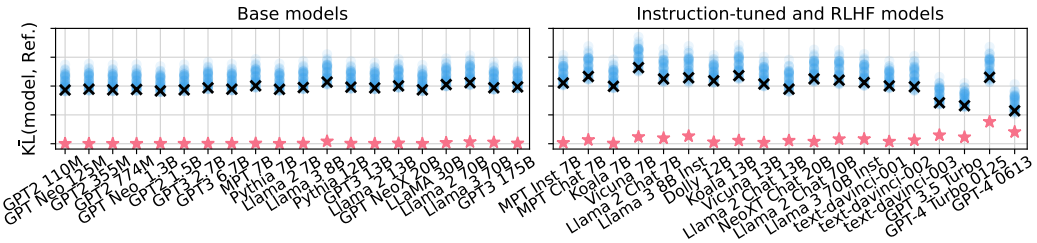


Figure 5: Divergence between adjusted model responses and different baselines: the overall U.S. census (✖), individual U.S. states (●), and a uniform baseline (★). Smaller means more similar. Model responses are by far more similar to the uniform baseline than to any human reference population.

4.2 Comparing model responses to the U.S. census

We now investigate the similarity of language models' adjusted responses to the census data. To do so, we consider the overall U.S. census population, as well as 50 census subgroups corresponding to every state in the United States. This leads to different human reference populations.

Inspired by the alignment measures proposed by Santurkar et al. [2023] and Durmus et al. [2023], we investigate the similarity of model responses to the census data by evaluating the average divergence across questions between model responses and the census statistics.[4] As we focus on categorical questions, we evaluate average KL divergence between each language model m and each reference population Ref, as follows:

$$\bar{\mathrm{KL}}(m, \mathrm{Ref}) = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{KL}(\bar{R}_q^m || \mathrm{Ref}_q).$$

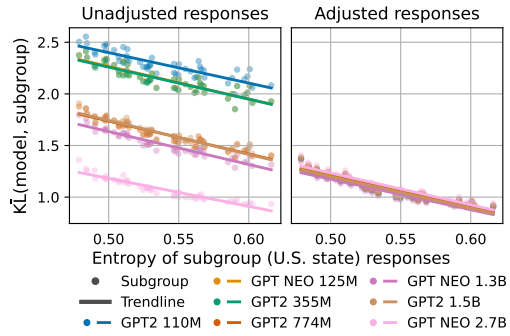
Results are depicted in Figure 5. For each model we plot the divergence to the census in black, the divergence to the different subgroups in blue, and the divergence to a uniform baseline with balanced responses in red. We observe that models are strikingly more similar to the uniform baseline than to any of the populations considered. For base models, this result is unsurprising, since in the previous section we established that base models' responses are essentially uniform after adjustment.

Looking at Figure 5 we find no consistent trend that instruction-tuning would move responses closer to the census, despite the increased deviation from uniform and the larger variations in entropy (recall Figure 4). Only for larger models the divergence seems to clearly decrease with instruction-tuning. However, all models' responses still remain significantly closer to the uniform baseline than to the U.S. census. For instance, for the GPT-4 model whose answers exhibit the highest similarity to the human reference populations, only 6 out of 25 questions (24%) are closer to the U.S. census than to the uniform baseline. Given these results, drawing conclusions about the relative alignment of models with subgroups is prone to resulting in brittle conclusions.

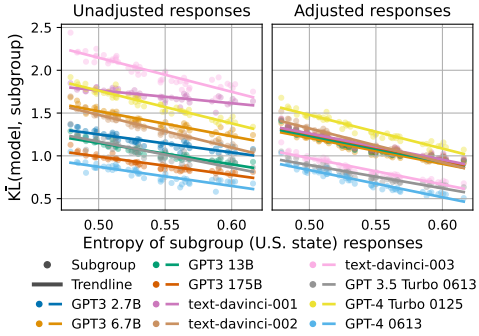
5 Implications for survey-based alignment metrics

Our findings add important context to previous works studying the alignment of language models with different human subpopulations. In particular, we highlighted the tendency of models towards balanced answers. Due to varying entropy in the responses of subgroups this leads to a strong correlation between model alignment and the reference population's entropy. The linear trend in Figure 6 visualizes this. For any given model, it consistently appears to be more "aligned" with the subpopulations exhibiting high entropy in their answers. Interestingly, we find that this trend also holds pre-adjustment, suggesting that the transformation of the response through randomized choice ordering is orthogonal to differentiating aspects of any specific population. In contrast, when comparing different models in Figure 6, we can see how adjustment has a large influence on their relative order. Differences across models that we see under naive prompting disappear after adjustment, which means that they should largely be attributed to systematic biases, rather than inherent properties of the model.

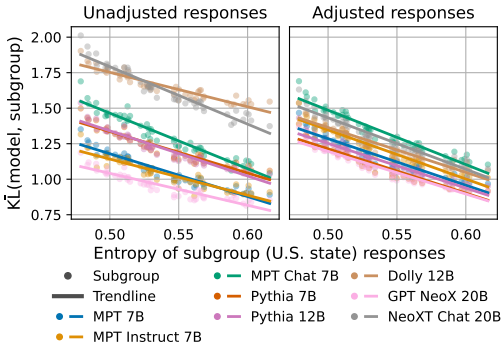
[4]Whereas Santurkar et al. [2023] use the Wasserstein distance to compare answer distributions, we use KL divergence since questions in the ACS are predominantly nominal, rather than ordinal.



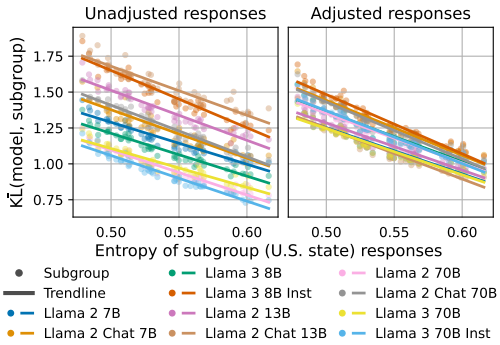
(a) GPT-2 and GPT-Neo.



(b) OpenAI's API models.



(c) MPT, Pythia, GPT-NeoX and its instruction variants.



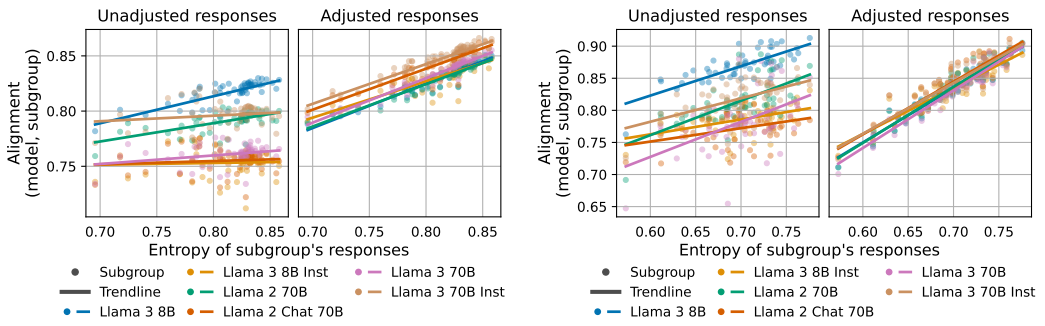
(d) Llama and its instruction and chat variants.

Figure 6: Alignment of models with different census subgroups. All models tend to exhibit similar relative alignment, and the divergence metric decreases with the entropy of the subgroups' responses.

Taken together our findings imply that the survey-derived alignment measure is more informative of differences in the reference populations rather than the language models is aims to evaluate. Model particularities, such as the pre-training data used, instruction tuning or the use of reinforcement learning with human feedback, seem to have little impact on which population is best represented.

5.1 Beyond the ACS

To inspect whether this trend changes with the content of the questions asked, we reproduce our experiments with additional surveys. We use the American Trends Panel (ATP) opinion surveys considered by Santurkar et al. [2023], and the Pew Research's Global Attitudes Surveys (GAS) and World Values Surveys (WVS) considered by Durmus et al. [2023]. These surveys encompass around 1500 questions and 60 U.S. demographic subgroups, and around 2300 questions and 60 national populations, respectively. We adopt the alignment metrics considered by the aforementioned works. We find that our insights gained from the ACS also hold for the ATP and GAS/WVS surveys. In particular, we similarly find a linear trend between the alignment metrics and subgroups' entropy of responses, in particular after adjustment, see Figure 7. Note here that alignment and divergence are negatively correlated by definition. Interestingly, this observation explains some of the findings in prior works. For example, Santurkar et al. [2023] find that "all the base models share striking similarities–e.g., being most aligned with lower income, moderate, and Protestant or Roman Catholic groups" and "our analysis [...] surfaces groups whose opinions are poorly reflected by current LLMs (e.g., 65+ and widowed individuals)". For the ATP surveys considered, low income, moderate, and Protestant/Catholic are precisely the demographic subgroups with responses closest to uniformly random among the income, political ideology, and religion demographic subgroups; whereas age 65+ and widowed are the demographic subgroups with responses furthest from uniform among the age and marital status demographic subgroups. Further, Santurkar et al. [2023] observe that RLHF can result in a "substantial shift [...] towards more liberal, educated, and wealthy [demographic



(a) ATP surveys.

(b) GAS/WVS surveys.

Figure 7: Alignment beyond ACS for selected models. We adopt the measures of Santurkar et al. [2023] and Durmus et al. [2023] on ATP and GAS/VVS opinion surveys. Again, the alignment between models and a given subpopulation correlates with the entropy of the subpopulations' responses.

groups]". Our results suggest that this could be an artifact of systematic biases. For the ATP surveys, we observe three outliers for which its alignment *before adjustment* is not correlated with the entropy of subgroup's responses: Llama 2 70B Chat and the two Llama 3 Instruct models. These are the models with largest pre-training compute considered. However, after adjustment, the alignment trends of Llama 2 70B Chat and the Llama 3 Instruct models are remarkably similar to that of their corresponding base models and all other LLMs.

6 Conclusion

We used a popular methodology to elicit LLMs' answer distributions to survey questions and closely examined the responses on the basis of the prime US demographic survey. We found that model responses are dominated by systematic ordering biases and do not exhibit the natural variations in entropy found in the human reference data collected by the US census. Even after adjusting for ordering biases, LLMs' responses still do not resemble those of human populations. Instead, they exhibit consistently high entropy, independent of the question asked. This holds true irrespective of model size or fine-tuning with human preferences.

These findings have important implications for insights gained from survey-derived alignment metrics. In particular, it explains why models of varying size all exhibit the same trend: they are most aligned with subgroups who happen to have balanced answers for the survey questions under consideration. For all models and surveys considered, alignment appears to be a proxy for the entropy of subgroups, rather than an inherent property of the model, or its training data.

We want to reiterate that our focus lies on questioning a popular methodology of eliciting survey responses from large language models using multiple choice prompting. At the example of this methodology our results highlight an important pitfall and suggest caution to expect robust insights when comparing such responses against those of human populations. The robustness and quality of an established survey does not seamlessly translate from the results obtained by surveying human populations to the logits output by LLMs. More research is urgently needed to design methodologies for getting insights into the inherent biases of LLMs and the population they might represent. Here public surveys and their accompanying data offer exciting potential and the could play an important role as a benchmarking tool for systematic evaluations of LLMs, see [Cruz et al., 2024] as an example. Although the use of survey data for LLM research has recently gained popularity, it still remains a widely under explored data source.

Acknowledgements

The authors would like to thank Frauke Kreuter and the Social Data Science and AI Lab at Ludwig-Maximilians-Universität Munich for inspiring discussions on an earlier version of this manuscript. Celestine Mendler-Dünner acknowledges financial support from the Hector Foundation.

References

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371, 2023.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, 2024.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: a suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, 2021.

James Brand, Ayelet Israeli, and Donald Ngwe. Using GPT for Market Research. *Harvard Business School Marketing Unit Working Paper No. 23-062*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020.

André F Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. *ArXiv preprint arXiv:2407.14614*, 2024.

Databricks. Dolly 12b, 2023. URL `https://github.com/databrickslabs/dolly`.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 2023.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 2021.

Ricardo Dominguez-Olmedo, Florian E Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. *ArXiv preprint arXiv:2407.07890*, 2024.

Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. Do personality tests generalize to large language models? In *NeurIPS Workshop on Socially Responsible Language Modelling Research*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *ArXiv preprint arXiv:2407.21783*, 2024.

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

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *Findings of the Association for Computational Linguistics*, 2023.

R.M. Groves, F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*. Wiley, 2009.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper*, 2023.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. CommunityLM: Probing Partisan Worldviews from Language Models. In *International Conference on Computational Linguistics*, 2022.

Junsol Kim and Byungkyu Lee. AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys. *ArXiv preprint arxiv:2305.09620*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Sanguk Lee, Tai-Quan Peng, Matthew H Goldberg, Seth A Rosenthal, John E Kotcher, Edward W Maibach, and Anthony Leiserowitz. Can large language models capture public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *ArXiv preprint arXiv:2311.00217*, 2023.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. Unqovering stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics*, pages 3475–3489, 2020.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *ArXiv preprint arxiv:2211.09110*, 2022.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 8086–8098, 2022.

Andrew Mao, Naveen Raman, Matthew Shu, Eric Li, Franklin Yang, and Jordan Boyd-Graber. Eliciting bias in question answering models through ambiguity. In *Workshop on Machine Reading for Question Answering*, pages 92–99, 2021.

Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. Do ais know what the most important issue is? using language models to code open-text social survey responses at scale. *SSRN Electronic Journal*, 2022.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, 2018.

MosaicML. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs, 2023. URL `www.mosaicml.com/blog/mpt-7b`.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: Measuring chatgpt political bias. *Available at SSRN 4372349*, 2023.

Laura K Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1):202–237, 2021.

OpenAI. Gpt-4 technical report. *ArXiv preprint arXiv:2303.08774*, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics*, pages 13387–13434, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023a.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023b.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. The Self-Perception and Political Biases of ChatGPT. *ArXiv preprint arXiv:2304.07333*, 2023.

Nathan E Sanders, Alex Ulinich, and Bruce Schneier. Demonstrations of the potential of ai-based political issue polling. *ArXiv preprint arXiv:2307.04781*, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *International Conference on Machine Learning*, 2023.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158, 2019.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *ArXiv preprint arXiv:2311.04076*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint arXiv:2307.09288*, 2023.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706, 2021.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024.

A Experimental details

We use the American Community Survey (ACS) Public Use Microdata Sample (PUMS) files made available by the U.S. Census Bureau.[5] The data itself is governed by the terms of use provided by the Census Bureau.[6] We download the data directly from the U.S. Census using the Folktables Python package [Ding et al., 2021]. We download the files corresponding to the year 2019.

We downloaded the publicly available language model weights from their respective official HuggingFace repositories. We run the models in an internal cluster. The total number of GPU hours needed to complete all experiments is approximately 1500 (NVIDIA A100). The budget spent querying the OpenAI models was approximately $200.

We open source the code to replicate all experiments.[7]

In addition, the repository contains notebooks to visualize the results of our investigations under different prompt ablations.

A.1 Survey questionnaire used

The exact questionnaire used in our experiments can be retrieved from our Github repository. We consider 25 questions from the 2019 ACS questionnaire corresponding to the following variables in the Public Use Microdata Sample: SEX, AGEP, HISP, RAC1P, NATIVITY, CIT, SCH, SCHL, LANX, ENG, HICOV, DEAR, DEYE, MAR, FER, GCL, MIL, WRK, ESR, JWTRNS, WKL, WKWN, WKHP, COW, PINCP. We take all questions as they appear in the ACS, with the exceptions:

- HISP: The ACS contains 5 answer choices corresponding to different Hispanic, Latino, and Spanish origins, and respondents are instructed to write down their origin if their origin is not among the choices provided. We instead provide two choices: “Yes” and “No".
- RAC1P: The ACS contains 15 answer choices, allows for selecting multiple choices, and respondents are instructed to write down their race if not among those in the multiple choice. The PUMS then provides up to 170 race codes (RAC2P and RAC3P). We instead present 9 choices, corresponding to the race codes of the RAC1P varible in the PUMS data dictionary.

Additionally, the variables ESR and COW are not directly associated with any single question in the ACS, but rather aggregate employment information. We formulate them as questions by taking the PUMS data dictionary’s variable and codes descriptions. Lastly, for the questions corresponding to the variables AGE, WKWN, WKHP, and PINCP, respondents are asked to write down an integer number. We convert such questions to multiple-choice via binning.

B Detailed experimental results

B.1 Model responses across questions before and after adjusting for A-bias

The results in this section complement Section 3, and pertain non-instruction-tuned language models. When surveying models without choice order randomization, we observe that the entropy of model responses tends to increase log-linearly with model size, often matching the entropy of the uniform distribution for the larger models. This trend is consistent across survey questions, irrespective of the question’s distribution over responses observed in the U.S. census (Figure 8).

B.2 A-bias of instruction-tuned models

The results in this section complement Section 3.1, and pertain instruction-tuned language models as well as language models fine-tuned with reinforcement learning with human feedback (RLHF). We observe that the strength of A-bias for these models, plotted in Figure 9, is comparable to that of base pre-trained models, plotted in Figure 3. This motivates the use of choice-order randomization in order to eliminate confounding due to labeling biases in models’ responses.

[5] https://www.census.gov/programs-surveys/acs/microdata.html

[6] https://www.census.gov/data/developers/about/terms-of-service.html

[7] https://github.com/socialfoundations/surveying-language-models

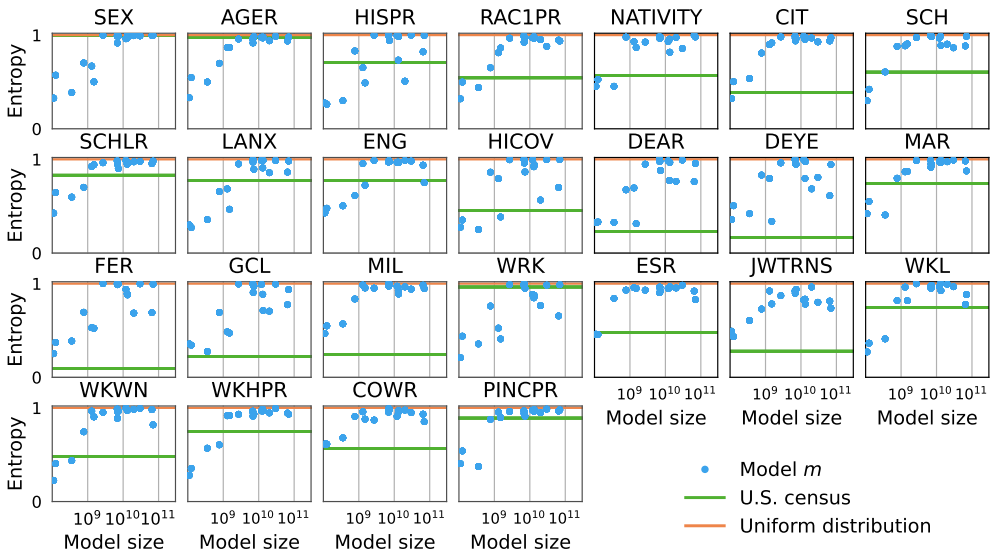


Figure 8: Normalized entropy of survey responses for individual questions (without adjustment).

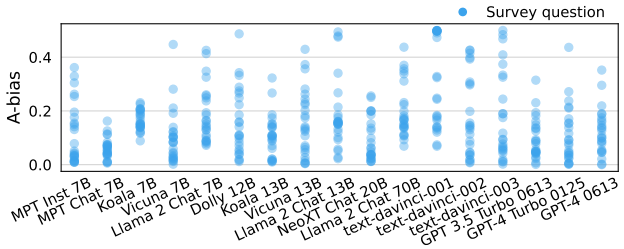


Figure 9: A-bias of instruction-tuned models.

B.3 Relative alignment across demographic subgroups

The results presented here complement those of Section 5. We plot the average KL divergence between each language model and each demopgrahic subpopulation (U.S. state) against the average entropy of the subgroup's responses. For readability, we split models into GPT-2 and GPT-Neo (Figure 6(a)), OpenAI's API models (Figure 6(b)), MPT, Pythia, GPT-NeoX and its instruction variants (Figure 6(c)), and LLaMA, Llama 2 and its instruction and chat variants (Figure 6(d)).

C Ordering bias: further experiments

We conduct additional randomization experiments pertaining to answer choice position and labeling bias, complimenting Section 3. We consider the GPT-2, GPT Neo, MPT, Pythia, and LLaMA models. The experiments follow a consistent setup:

1. We randomize both the order in which choices are presented and the label (i.e., letter) assigned to each answer choice. For example, for the "sex" question, the possible combinations are "A. Male B. Female", "A. Female B. Male", "B. Male A. Female", and "B. Female A. Male". Note that in the experiments presented in Section 3.1 we only randomized over the order in which choices are presented (i.e., the "A" choice was always presented first).
2. We compute the output distribution over responses for choice position (the probability assigned to the first, second, etc., answer choice presented) and letter assignment (the probability assigned to the answer choice assigned "A", "B", etc.).

For each model and survey question, we estimate the expected distribution over responses for both choice position and letter assignment by collecting 3,000 responses (step 2) under different

randomizations of choice position and letter assignment (step 1). A model with no position and labeling biases would assign the same probability distribution to answer choices (e.g., "male" and "female") regardless of position or letter assignment, and therefore the expected distributions over position (e.g., selecting the first choice) and letter assignment (e.g., selecting "A") would be uniform.

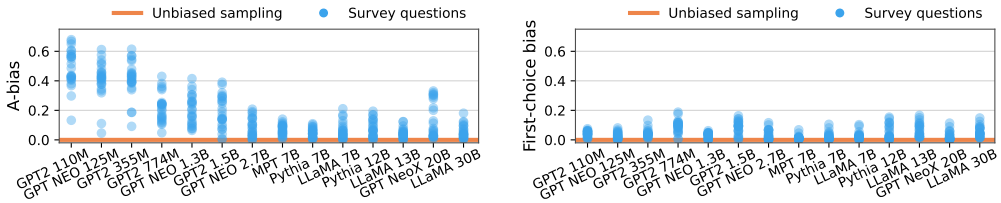
C.1 Disentangling ordering bias into positioning bias and labeling bias

We perform chi-square tests to determine whether language models' output responses distributions over position and letter assignment significantly deviate from the uniform distribution (i.e., if there exists statistically significant bias in position or letter assignment). Since we collect 3,000 response distributions under randomized choice position and letter assignment, we ensure a high test power ($\geq$ 0.98) in detecting small effect sizes (0.1) at a significance level of 0.05.

We find that models exhibit significant positioning and labelling for most survey questions, see Figure 10. We observe that labelling is more prevalent that positioning bias. While both tend to decrease with model size, order bias decreases more significantly with model size, whereas labeling bias tends to be very prevalent across all model sizes. In Figure 11 we plot both the strength of A-bias and first-choice bias across survey questions. The strength of A-bias tends to be greater than that of first-choice bias, particularly for the smaller models.

(a) Significant ordering bias. (b) Significant label bias.

Figure 10: All models exhibit statistically significant letter and ordering bias for most survey questions.



(a) A-bias (b) First choice bias

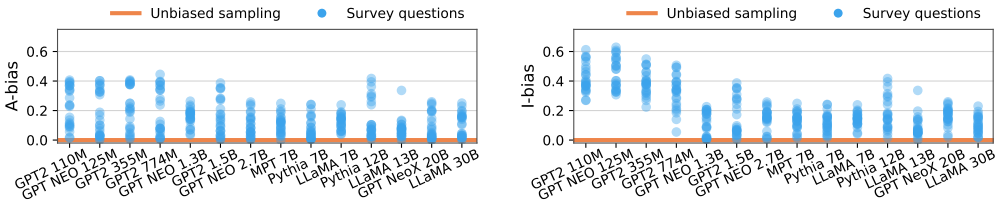
Figure 11: Models, particularly those with less than a few billion parameters, tend to exhibit stronger A-bias than first-choice bias.

C.2 I-bias

We hypothesize that A-bias is prevalent because the single character "A" is relatively frequent as the starting word of a sentence in written English. We test this hypothesis by replacing the character "B" with "I" when presenting the survey questions, since the character "I" is even more frequent as the starting word of a sentence in written English. We randomize over choice ordering and label assignment as in the previous evaluation. We find that, when presenting both "A" and "I", small models then exhibit I-bias rather than A-bias (Figure 12), supporting our initial hypothesis.

C.3 Using letters with similar frequency in written English

Motivated by the I-bias experiment, we now examine whether labeling bias can be mitigated by using letters that have similar frequency in written English. Therefore, instead of assigning to choices the



(a) A-bias in the "A", "I" randomization experiment. (b) I-bias in the "A", "I" randomization experiment.

Figure 12: When both "A" and "I" are present, small models exhibit I-bias rather than A-bias.

(a) Significant labelling bias.

(b) Significant ordering bias.

Figure 13: "R", "S", "N", etc. randomization experiment. All models, irrespective of size, exhibit statistically significant letter and positioning bias for most survey questions.

labels "A", "B", etc. we assign the following labels: "R", "S", "N", "L", "O", "T", "M", "P", "W", "U", "Y", "V". We find that, compared to the "A", "B", etc. randomization experiment, the percentage of questions for which models exhibit significant labeling bias somewhat decreases (Figure 13). However, models tend to exhibit substantially more position bias. This indicates that, in the absence of a label that provides a strong signal (e.g., "A" or "I"), models tend to exhibit significantly higher choice-ordering bias, irrespective of model size.

D Prompt ablations

We reproduce our experiments using different prompts to query the model. Due to the cost of querying OpenAI's models, we only perform these ablations for models with publicly available weights. The notebooks with all figures can be retrieved from our Github repository. [8]

Overall, the prompt ablation results are very consistent with the findings presented in the main text of the paper. In the following we provide an overview over the different ablations performed.

D.1 System rompt used for GPT-3.5 and GPT-4

When querying GPT-3.5, GPT-4, and GPT-4 Turbo, we use the system prompt `Please respond with a single letter.`, as otherwise for most questions none of the top-5 logits correspond to answer choice labels (e.g., "A", "B"). Note that this problematic arises due to the fact that the OpenAI API only allows access to the top 5 logits. We adapt the system prompt used by Dorner et al. [2023] in the context of surveying GPT-4 with standarized personality tests.

D.2 Individual survey questions

First, we use different styles to prompt individual survey questions. We enumerate the prompt styles as (P1)-(P8).

Additional context. We first explore whether including additional context signaling that the questions presented are from the American Community Survey, or that they are to be answered by

[8] https://github.com/socialfoundations/surveying-language-models/blob/main/prompt-ablations

U.S. households. Keeping identical survey questions, we append at the start of the prompt one of the following sentences:

(P1) `Bellow is a question from the American Community Survey.`

(P2) `Answer the following question from the American Community Survey.`

(P3) `Answer the following question as if you lived at a household in the United States.`

Asking questions in the second person. We change the framing of the questions.

(P4) We modify the survey questionnaire such that questions are formulated in the second person rather than the third person (e.g., "What is your sex?" instead of "What is this person's sex?").

Including instructions. Following the prompt ablation of Santurkar et al. [2023], we append at the start of the prompt one of the following instructions:

(P5) `Please read the following multiple-choice question carefully and select ONE of the listed options.`

(P6) `Please read the multiple-choice question below carefully and select ONE of the listed options. Here is an example of the format:\nQuestion: Question 1\nA. Option 1\nB. Option 2\n C. Option 3\nAnswer: C`

Chat-style prompt. We consider the prompt used by Durmus et al. [2023]:

(P7) `Human: {question}\nHere are the options:\n{options}\n Assistant: If had to select one of the options, my answer would be`

Interview-style prompt. We consider the prompt used by Argyle et al. [2023]:

(P8) `Interviewer: {question}\n{options}\nMe:`

D.3 Sequential generation

We use different prompts to integrate a model's previous responses when prompting subsequent survey questions. Instead of summarizing previous responses using bullet points as in Section 5, we keep previous questions and answers in-context.

Question answering. Keeping questions and answers in-context resembles the typical few-shot Q&A setting. For instance, prompting for the third question in the questionnaire corresponds to

```
Question: {question 1}\n{options 1}\nAnswer:{answer 1}\n
Question: {question 2}\n{options 2}\nAnswer:{answer 2}\n
Question: {question 3}\n{options 3}\nAnswer:
```

Interview-style prompt. We consider the prompting style used by Argyle et al. [2023]. For instance, prompting for the third question in the questionnaire corresponds to

```
Interviewer: {question 1}\n{options 1}\nMe:{answer 1}\n
Interviewer: {question 2}\n{options 2}\nMe:{answer 2}\n
Interviewer: {question 3}\n{options 3}\nMe:
```

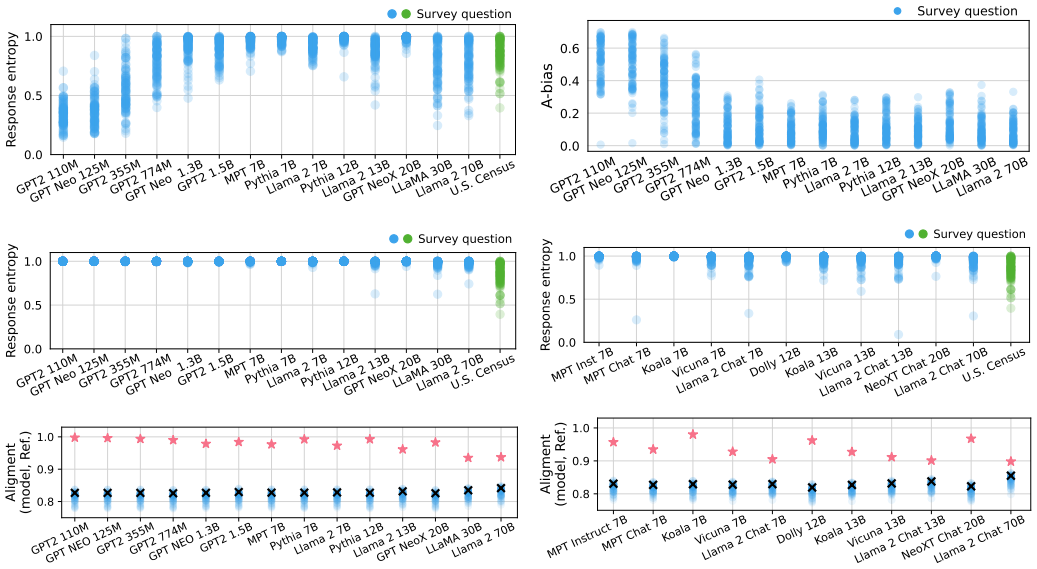


Figure 14: Reproduction of the experiments in Sections 3 and 4 for the ATP surveys.

E Results for ATP, GAS, WVS, and ANES surveys

We reproduce the experiments of Sections 3 and 4 using the ATP, and GAS/WVS used by Santurkar et al. [2023] and Durmus et al. [2023], where questions are presented individually of one another. We additionally reproduce the experiments of Section 5 using the 2016 ANES questionnaire considered by Argyle et al. [2023], where questions are presented in sequence. We do not consider OpenAI's models as the cost to reproduce the experiments via the OpenAI API exceeds our budget. We obtain very similar results to those of the ACS presented in the main text of the paper. The notebooks with all figures can be retrieved from our Github repository. [9]

E.1 ATP surveys

We obtain the ATP survey questions and their corresponding human responses from the OpinionsQA repository.[10] We present all answer choices when querying the models, but exclude the answer choices corresponding to refusals from our analysis similarly to Santurkar et al. [2023]. When comparing the similarity of models' responses to different demographic subgroups, we use the demographic subgroups and the alignment metric considered by Santurkar et al. [2023]. For such metric, higher values of alignment indicate that models' responses are more similar to the reference demographic group. We find that all models are more "aligned" with the uniformly random baseline than with any of the demographic subgroups, see Figure 14.

E.2 GAS and WVS surveys

We obtain the ATP survey questions and their corresponding human responses from the GlobalOpinionsQA repository.[11] When comparing the similarity of models' responses to the population-level survey responses of different countries, we use the countries and the similarity metric considered by Durmus et al. [2023]. We find that all models produce survey responses that are more similar to those of the uniformly random baseline than to those of any of the demographic subgroups, see Figure 15.

[9] https://github.com/socialfoundations/surveying-language-models

[10] `https://github.com/tatsu-lab/opinions_qa`

[11] `https://huggingface.co/datasets/Anthropic/llm_global_opinions`

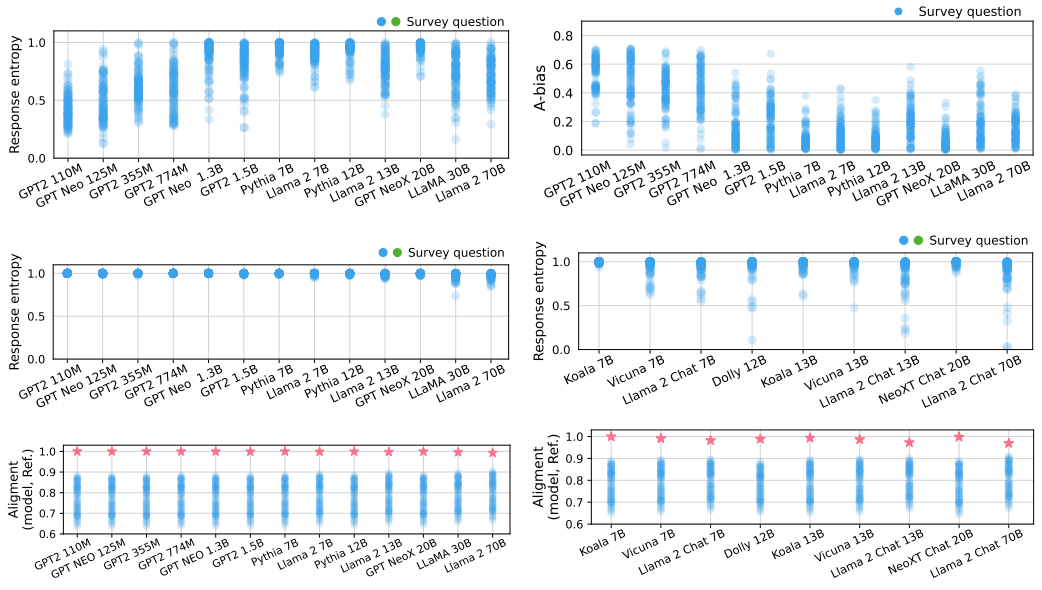


Figure 15: Reproduction of the experiments in Sections 3 and 4 for the GAS/WVS surveys.

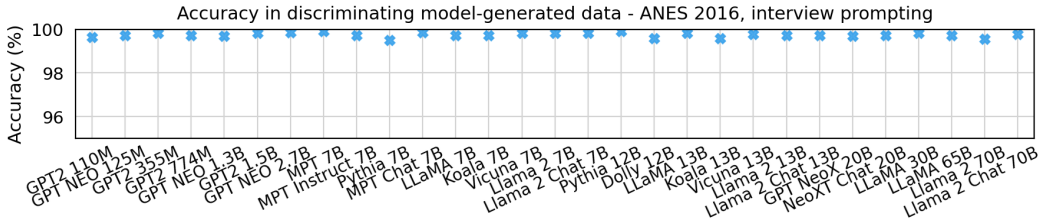


Figure 16: The discriminator test performed on datasets generated using the 2016 ANES survey questionnaire (with choice randomization).

E.3 Relative alignment for ATP and GAS/WVS surveys

We consider the alignment measures proposed by Santurkar et al. [2023] and Durmus et al. [2023] on ATP and GAS/VVS opinion surveys for the largest base / instruct models considered. We find that, similarly to our observations for the ACS, the alignment between models and a given subpopulation is highly correlated with the entropy of the subpopulations' responses.

E.4 ANES survey

We present questions in the multiple-choice format described in Section 2, using the `Interviewer:`, `Me:` prompt style described by Argyle et al. [2023]. We retrieve the 2016 ANES data from the official website[12], and process it such that it matches in form the questionnaire designed by Argyle et al. [2023]. We find that the trained classifiers can discriminate between the model-generated data and the ANES data with very high accuracy ($\geq$99%), see Figure 16.

F Sequential sampling of responses

Motivated by recent findings of Argyle et al. [2023] we conducted an additional investigation where we seek to fill entire ACS questionnaires in a sequential manner, in order to generate for each language model a synthetic dataset of responses. This data emulates in form the ACS dataset collected by the

[12] `https://electionstudies.org/data-center/2016-time-series-study/`

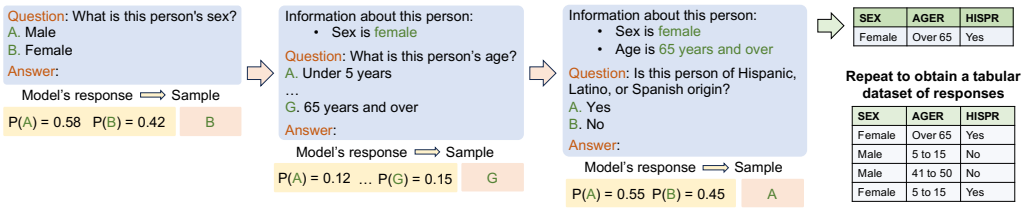


Figure 17: Methodology and prompt template used to sequentially sample models' responses to entire survey questionnaires. We provide the answers to previous question in context when prompting subsequent questions. The output is a tabular dataset of responses.

U.S. Census Bureau. We then study the extent to which such synthetic datasets resemble the ACS dataset.

F.1 Methodology

We present survey questions in the same order as in the ACS questionnaire. When querying a model to answer survey question q, we include a summary of the $q-1$ previously sampled answers in context. [13] We then sample from the model's output probability distribution over answers, and continue to the next question. We illustrate this sequential process in Figure 17. We refer to Appendix D.3 for results collected with different variations of how a model's previous answers are integrated into the prompt. We find our results to be robust to these prompt variations.

For each language model we sample N=100,000 model-generated responses to the ACS. Due to the cost of querying OpenAI's models, we only survey GPT-4 and sample $N = 500$ responses. As a result, we generate for each language model a tabular dataset similar in form to the ACS data, with N rows corresponding to each filled questionnaire and 25 columns corresponding to each question.

F.2 The discriminator test

We investigate whether the model-generated datasets resemble the U.S. census data by constructing a binary prediction task aiming to discriminate synthetic responses from census responses. Intuitively, if the two datasets were very dissimilar, then a classifier would be able to achieve high accuracy. Formally, let $\mathcal{F}$ be class of binary prediction functions mapping each data point (i.e., a row in the tabular dataset) to $\{0, 1\}$, then the accuracy of the best $f \in \mathcal{F}$ on the discriminator task provides a lower bound on the total variation (TV) distance between the two empirical data distributions.

Hence, we train a predictor f to discriminate between the model-generated data and the census data in order to obtain an empirical lower bound on the distance between the two datasets. Specifically, we concatenate to each model-generated dataset a random sample of N individuals from the ACS census data, and introduce a binary label indicating whether each row of the concatenated dataset was model-generated or not. We then train an XGBoost classifier in this binary prediction task. As an additional point of reference, we also consider the accuracy in discriminating between the census data of any given U.S. state and an equally-sized sample of the ACS data of all other U.S. states.

We report mean test accuracy in Figure 18. We consider 100 different random seeds. We find that the trained classifiers can differentiate between model-generated data and census data with very high accuracy ($> 90\%$) in all cases. Therefore, the empirical distributions corresponding to the model-generated data and the census data have TV distance larger than 0.9. These stark results indicate that data generated by sequentially prompting language models with the ACS survey questionnaire bears little similarity with the data collected by surveying the U.S. population.

F.3 Contrast with silicon samples

Argyle et al. [2023] propose "silicon sampling", a methodology to produce synthetic survey respondents using LLMs by conditioning on actual survey respondents. They focus on a subset of

[13] The maximum number of tokens in a filled questionnaire was less than 1024 tokens in all cases, thus fitting entirely within the context window of all surveyed models.

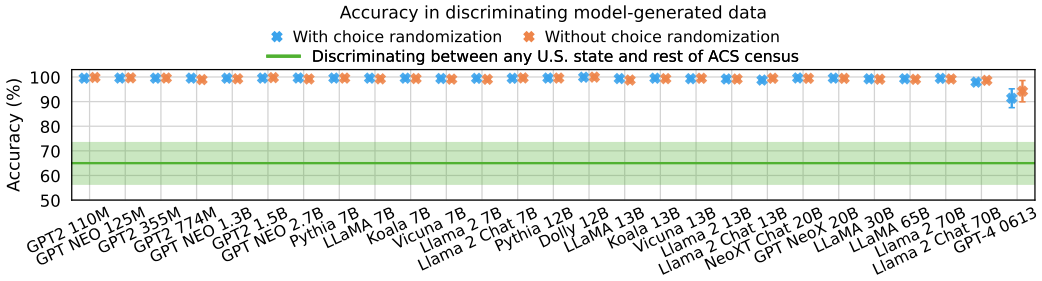


Figure 18: Accuracy of the discriminator test. For all language models, it is possible to discriminate with very high accuracy between the ACS census data and model-generated data, (✖) before adjustment and (✖) after adjustment. We contrast this against the accuracy value of discriminating between the ACS data of any given U.S. state and the rest of the ACS census data (–).

12 questions from the 2016 American National Election Studies (ANES) survey. For every human respondent, they construct a corresponding "silicon individual" by querying GPT-3 to predict the ANES respondent's answer to each survey question given the respondent's answers to all other questions. Their results indicate that, for the 2016 ANES survey, GPT-3 can be a fairly calibrated predictor of an individual's answer to some survey question conditioned on the respondent's answers to all other survey questions.[14]

However, Argyle et al. [2023] emphasize that important insights can be gained by emulating the survey responses of human populations "prior to or in the absence of human data". In this work we have considered precisely the setting where models' responses are obtained in the absence of human data.[15] To investigate how our findings transfer to the ANES, we reproduce the experiments of Section F using the 2016 ANES survey questionnaire considered by Argyle et al. [2023] and their "interview-style" prompt. We apply the discriminator test, and find that the trained classifiers can discriminate between the model-generated data and the ANES data with accuracy > 99 % (see Appendix E), indicating that models' responses are markedly different to those in the ANES data.

Thus, the fact that models may perform reasonably well at feature imputation tasks (e.g., *predicting* an individual's answer to some question given their answers to *all other questions*) does not imply that models can *generate* synthetic respondents that resemble the responses obtained by surveying human populations. This suggests caution when using LLMs to emulate human populations at present time, in particular in the absence of human data.

[14] Lee et al. [2023] and Sanders et al. [2023] study imputation tasks similar to those of Argyle et al. [2023], and find that LLMs are not calibrated predictors for a variety of such tasks.

[15] Conceptually, to generate each synthetic individual, we start with a blank survey questionnaire and prompt the LLM to sequentially fill the entire questionnaire. Whereas we only prompt LLMs with their own responses (i.e., to previous survey questions), Argyle et al. [2023] prompt the model only with actual human responses from the 2016 ANES data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction directly refer to the experiments and results detailed in the paper's main body.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, see Section 2 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix A and the code release.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main figures contain exact measures. We conduct significance tests on Section B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, see Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We argue throughout the paper that current evaluation practices might result in misleading claims regarding what subgroups current models best represent.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any models. We release models' survey responses, which pose no risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the authors of the models considered, as well as the sources of the surveys considered.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.