QUESTIONING THE SURVEY RESPONSES OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models increase in capability, researchers have started to conduct surveys of all kinds on these models in order to investigate the population represented by their responses. In this work, we critically examine language models' survey responses on the basis of the well-established American Community Survey by the U.S. Census Bureau and investigate whether they elicit a faithful representation of any human population. Using a de-facto standard multiple-choice prompting technique and evaluating 39 different language models using systematic experiments, we establish two dominant patterns: First, models' responses are governed by ordering and labeling biases, leading to variations across models that do not persist after adjusting for systematic biases. Second, models' responses do not contain the entropy variations and statistical signals typically found in human populations, but strongly tend towards uniform answers. As a result, models' relative alignment with different demographic subgroups can be predicted from the subgroups' entropy, irrespective of the model's training data or training strategy. Our findings add important context to recent works that investigate the alignment of language models with demographic subgroups.

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). 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' alignment (Santurkar et al., 2023; Durmus et al., 2023).[1]

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 population they represent. The quality of survey data hinges on the validity and robustness of the conclusions that can be drawn from it.

In this work we investigate the survey responses of LLMs on the basis of the American Community Survey (ACS), the premier demographic survey conducted by the U.S. Census Bureau. We prompt 39 language models of varying size with questions from the ACS, and based on the collected data, we investigate whether models' responses elicit a faithful representations of any human population.

2 SURVEYING LANGUAGE MODELS

We employ the de-facto standard methodology to survey language models introduced by Santurkar et al. (2023). 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. We determine the event probabilities of R_q^m as follows:

[1] For a detailed discussion of related work, see Appendix C.4.

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]
2. We query the language model m with the input prompt and obtain its 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.

Due to space constraints, we only include in the main text results for GPT-2, GPT-3, and GPT-4. See Appendix C for the complete results of all surveyed language models.

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, with few exceptions. See Appendix B.1 for further information.

Reference data and 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 (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. We use U_q to denote the uniform distribution. We use normalized *entropy* to measure the degree of variation in models' responses, and we use the *Kullback–Leibler (KL) divergence* to measure the similarity between two answer distributions.

Randomized choice ordering. For several investigations we survey models under randomized choice ordering. This is an established methodology in survey research to adjust for ordering biases (Groves et al., 2009). For a given question q, we prompt models with all possible permutations of the answer choice ordering. Choice labels are presented in alphabetical order in all cases. 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 under choice ordering randomization. This distinction serves to decouple a model's tendency towards picking a particular answer from its tendency towards picking a particular choice label. See Figure 4 in Appendix B.2 for an illustration of the methodology.

3 Inspecting models' survey responses

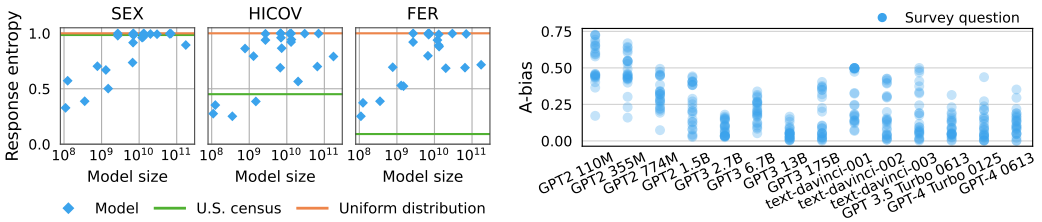
We start by surveying the base pre-trained models. For a first investigation, we consider the normalized entropy of models' responses to the SEX, HICOV, and FER questions, which inquiry about the person's sex, whether they are currently covered by any health insurance plan, and whether they gave birth in the past 12 months, respectively. When surveying the U.S. population, these three questions elicit responses with very different entropy (e.g., sex is relatively balanced but most people do not give birth in any given year). In contrast, as shown in Figure 1(a), the entropy of models' responses to these three questions are surprisingly similar. In particular, we find that response entropy tends to increase log-linearly with model size. This trends holds across all ACS questions, see Appendix C.1.

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 variations across models might arise mostly due to systematic biases.

3.1 Testing for systematic biases: A-bias

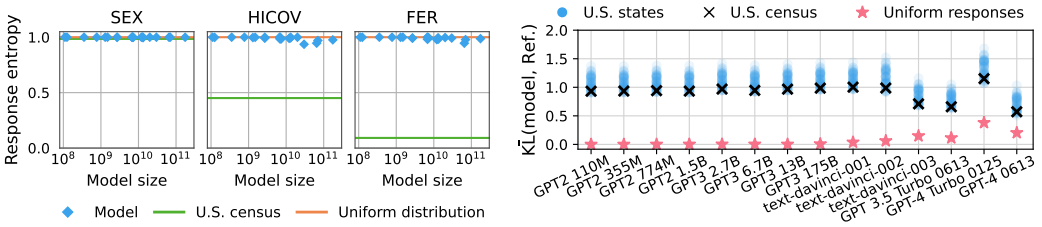
It is well-known that language models' most likely answer to multiple-choice questions can change depending on seemingly minor factors such as the ordering of the answer choices (Robinson & Wingate, 2023). Instead, we seek to measure the extent to which changes in choice ordering alter a model's output *distribution over answers*. We start by measuring *A-bias*: the tendency of a model towards picking the answer choice labeled "A". For an unbiased model that outputs the same answer distribution irrespective of choice ordering, the expected choice distribution $\bar{O}_q^m$ under randomized

[2]The chosen style of prompt is standard for question answering tasks (Hendrycks et al., 2021), and used in OpinionQA (Santurkar et al., 2023). We perform several prompt ablations, including the prompt variations used by Argyle et al. (2022), Santurkar et al. (2023) and Durmus et al. (2023), see Appendix E.



(a) Entropy of base models' responses. (b) A-bias of models' responses.

Figure 1: Entropy and A-bias of models' responses. (a) Models' variation in entropy across questions is much smaller than that of the census data. (b) All models suffer from substantial A-bias.



(a) Entropy of base models' *adjusted* responses. (b) Divergence to different reference populations.

Figure 2: After adjustment, (a) Base models tend towards uniform answers, and (b) Models' responses are more similar to the uniform baseline than to any of the populations considered.

choice ordering would match precisely the uniform distribution. We define A-bias as a models' deviation from this unbiased baseline, that is, $\text{Abias}_q^m := \left|\text{P}(\bar{O}_q^m = \text{“}A\text{”}) - 1/k_q\right|$.

We measure A-bias for each question q and model m. Results are illustrated in Figure 1(b). We 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.

We investigate other types of labelling and position bias (e.g., last-choice bias) in Appendix D. 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 LLMs' exhibit substantial response biases when prompted with multiple-choice survey questions, and that these biases are generally not human-like.

4 Controlling for A-bias

To eliminate confounding due to labeling biases, we survey models under randomized choice ordering. We refer to the expected response under choice order randomization as the *adjusted* response.

In Figure 2(a) we plot the normalized entropy of base models' responses after adjustment. 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, base models' survey responses are approximately uniform irrespective of model size or survey question asked. This validates our initial hypothesis that, without adjustment, variations in responses across models arise predominantly due to systematic biases such as A-bias.

4.1 Comparing model responses to the U.S. census

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. We evaluate the average KL divergence across questions between adjusted model responses and those of different reference pop-

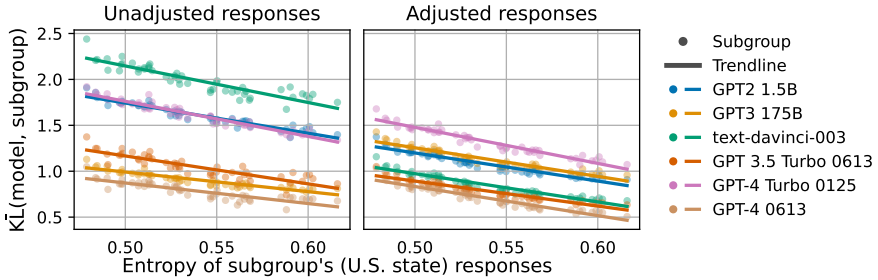


Figure 3: Alignment of models with different census subgroups strongly correlated with the subgroups' entropy of responses, both before and after adjusting for choice ordering biases.

ulations, that is, $\bar{\mathrm{KL}}(m, \mathrm{Ref}) = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{KL}(\bar{R}_q^m || \mathrm{Ref}_q)$. We consider the responses of the overall U.S. census, as well as 50 census subgroups corresponding to every state in the United States.

Results are depicted in Figure 2(b). We observe that models are strikingly more similar to the uniform baseline than to any of the populations considered. Furthermore, due to the models' tendency towards balanced answers, we observe a strong correlation between a models' similarity (or alignment) with a subgroup and the subgroup's entropy, as shown in Figure 3 (see Appendix C.4 for all models). Interestingly, this trend also consistently holds pre-adjustment.

Our results indicate that survey-derived alignment measures may be more informative of the reference populations rather than the language models they aim to evaluate. Particularities, such as model size, the training data used, or the demographics of the annotators used for fine-tuning with human feedback seem to have little impact on which population is best represented, and models consistently appear to be more "aligned" with the subpopulations exhibiting high entropy in their answers.

Beyond the ACS. To inspect whether this trend changes with the content of the questions asked, we reproduce our experiments with 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 the insights gained from the ACS also hold for the ATP and GAS/WVS surveys, see Appendix F.

In particular, we similarly find a linear trend between the alignment metrics considered and subgroups' entropy of responses, see Figure 16 in Appendix F.3. 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, low income, moderate, and Protestant/Catholic are the demographic subgroups with responses closest to uniformly random among the income, political ideology, and religion demographic subgroups; whereas age 65+ and widowed have responses furthest from uniform among the age and marital status subgroups.

5 Conclusion

We examined the survey responses of LLMs on the basis of the prime demographic survey of the United States. To do so, we leveraged a popular methodology to elicit LLMs' answer distributions to survey questions. We found that model responses are dominated by systematic ordering biases and do not exhibit the natural variations in entropy found in the census data. Even after adjusting for ordering biases, LLMs' responses still do not resemble those of human populations, irrespective of model size or fine-tuning with human preferences, but rather trend towards uniform responses.

Taken together, our findings caution to expect robust insights when comparing LLMs' responses against those of human populations. In our study we could not find any indication that LLMs elicit faithful a representation of any human population. Thus, the validity of surveys as an instrument to measure general properties of LLMs, such as alignment, is unclear at present time. 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 large language models.

REFERENCES

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 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*, pp. 337–371. PMLR, 2023.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*, 2022.

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. *arxiv prepring arxiv:2304.01373*, 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*, pp. 1877–1901, 2020.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

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.

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.

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.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. Blog post, 2023. URL `https://bair.berkeley.edu/blog/2023/04/03/koala/`.

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 *Proceedings of the 29th International Conference on Computational Linguistics*, 2022.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

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*, pp. 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 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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 *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pp. 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk's language. In *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pp. 589–612. Association for Computational Linguistics (ACL), 2022.

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, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, 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. *arXiv preprint arXiv:2212.09251*, 2022.

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 *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

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

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.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, 2020.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 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, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

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*, 2023b.

Tianping Zhang, Shaowen Wang, Shuicheng Yan, Jian Li, and Qian Liu. Generative table pre-training empowers models for tabular prediction. *arXiv preprint arXiv:2305.09696*, 2023.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arxiv preprint arxiv:2305.03514*, 2023.

A Related work

Despite the syntactical similarities, evaluating LLMs on the basis of their survey responses differs from 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 (Radford et al., 2019). Alternatively, models' most likely response to questions that lack a clear answer (e.g., "Angela and Patrick are sitting together. Who is an entrepreneur?") have been used to investigate various biases of LLMs (Li et al., 2020; Mao et al., 2021; Perez et al., 2022; Abid et al., 2021; Jiang et al., 2022).

When evaluating LLMs on the basis of survey questions, it is not models' most likely completion that is studied, but rather models' probability distribution over various answer choices. 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. Durmus et al. (2023) employ a similar methodology but instead consider transnational opinion surveys. We adopt this popular methodology to investigate the properties of models' answer distributions on the basis of a well-established demographic survey, beyond measuring the relative similarity of models' responses to different human populations.

In addition to asking questions individually, we also prompt models to complete entire survey questionnaires. We present questions in a sequential manner, keeping a model's previous answers in context when prompting the model to answer subsequent questions. This methodology resembles prior work by Hartmann et al. (2023); Rutinowski et al. (2023); Motoki et al. (2023) who sequentially prompt ChatGPT to answer entire political compass or voting advice questionnaires. But instead of aggregating answers into a political affinity score, our focus is on examining whether models' responses resemble those of human populations.

There is an emerging body of research that integrates LLMs into computational social science (Ziems et al., 2023). This includes tasks such as taxonomic labeling, where language models are employed for tasks such as opinion prediction (Kim & 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., 2022; 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.

Previous works have identified that the performance of language models in QA tasks can vary significantly depending on the input prompt (Shin et al., 2020; Jiang et al., 2020; Mishra et al., 2022), such as the order in which few-shot examples are presented (Zhang et al., 2023; Lu et al., 2022). Robinson & Wingate (2023) identify that in zero-short multiple-choice QA, a model's most likely answer can change depending on the order in which answer choices are presented. While we also study models' sensitivity to answer choice ordering, we instead study the extent to which changes in choice ordering affect a model's output distribution over answers. The concurrent work of Tjuatja et al. (2023) contrasts the response biases of LLMs to multiple-choice survey questions against those of humans, and finds that LLMs' response biases are generally not human-like.

B EXPERIMENTAL DETAILS

We use the American Community Survey (ACS) Public Use Microdata Sample (PUMS) files made available by the U.S. Census Bureau.[3] The data itself is governed by the terms of use provided by the Census Bureau.[4] 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 consider the base models GPT-2 (Radford et al., 2019), GPT-Neo (Black et al., 2021), Pythia (Biderman et al., 2023), MPT (MosaicML, 2023), LLaMA (Touvron et al., 2023a), Llama 2 (Touvron et al., 2023b), 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), the Vicuna and Koala fine-tunes of LLaMA 7B and 13B (Geng et al., 2023; Chiang et al., 2023), Llama 2 Chat (Touvron et al., 2023b), GPT-3.5, GPT-4 (OpenAI, 2023), and the text-davinci variants of GPT-3 (Ouyang et al., 2022).

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.

For the code and data to reproduce all experiments and plots, refer to

https://drive.google.com/drive/folders/1HEPo54-G7fthX7JEyws0MuvJFBk8x7Tt?usp=sharing

The supplementary material additionally contains notebooks to visualize the results of our investigations for different prompt ablations.

B.1 SURVEY QUESTIONNAIRE USED

The exact questionnaire used in our experiments can be retrieved from

https://drive.google.com/drive/folders/1HEPo54-G7fthX7JEyws0MuvJFBk8x7Tt?usp=sharing

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.2 ADDITIONAL DETAILS ON ADJUSTMENT

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

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

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

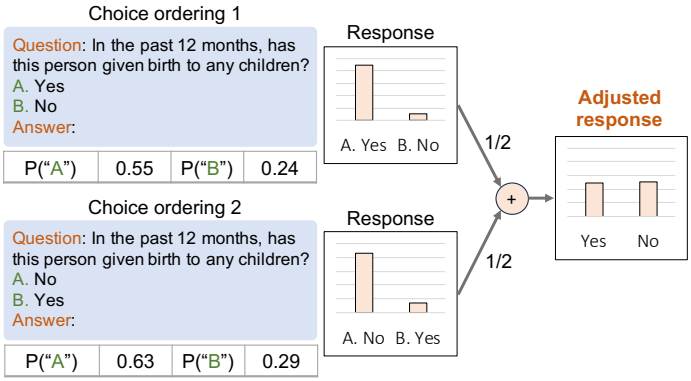


Figure 4: Adjustment methodology. We compute the expected survey response under all possible choice ordering permutations.

remaining probability mass and the smallest observed probability, following the methodology of Santurkar et al. (2023). See Figure 4 for an illustration of the adjustment methodology.

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.

B.3 System prompt 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.

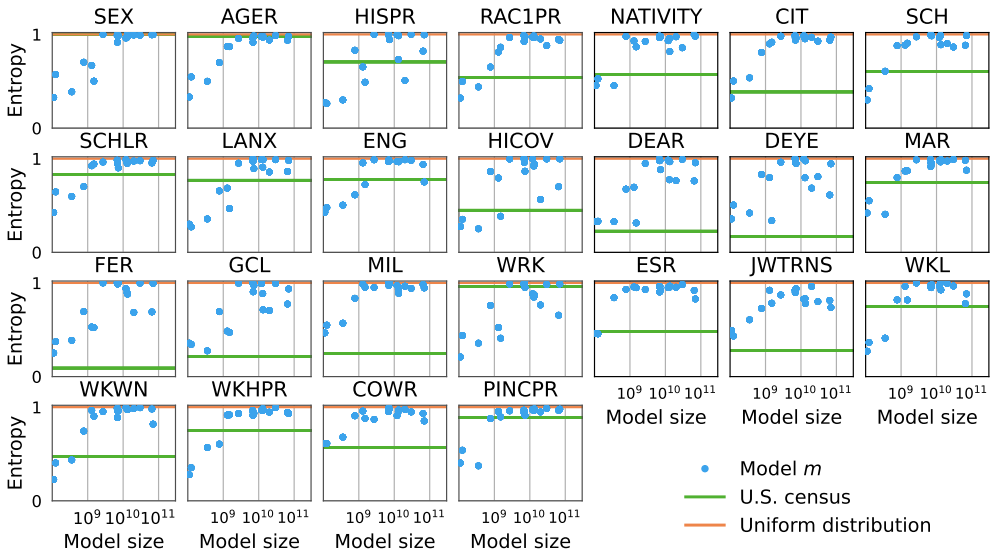


Figure 5: Per-variable normalized entropy of survey responses (without adjustment).

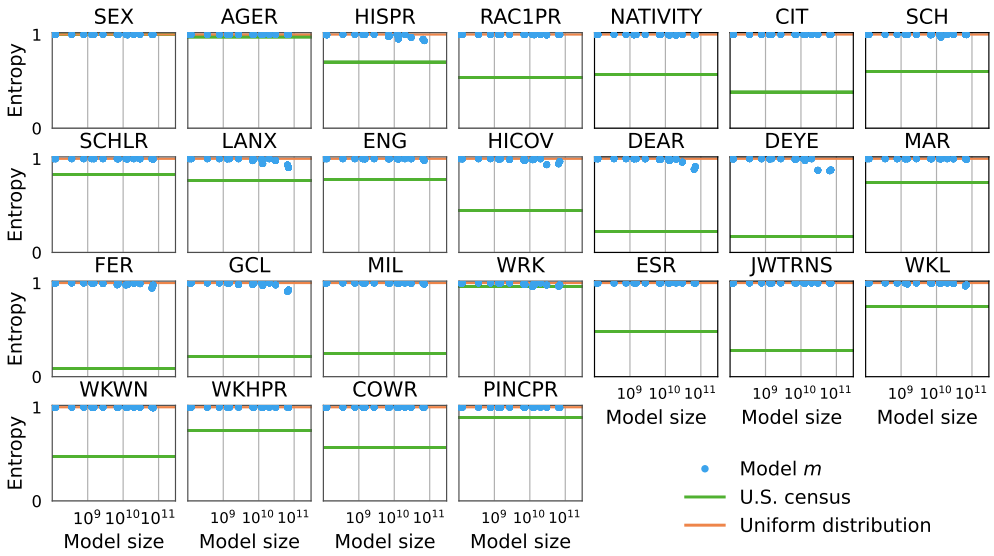


Figure 6: Per-variable normalized entropy of survey responses (with adjustment).

C Detailed Experimental Results

C.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 5). After adjusting for choice ordering bias via randomized choice orderings, language models' survey responses are highly entropic and strongly trend towards the uniform distribution (Figure 6).

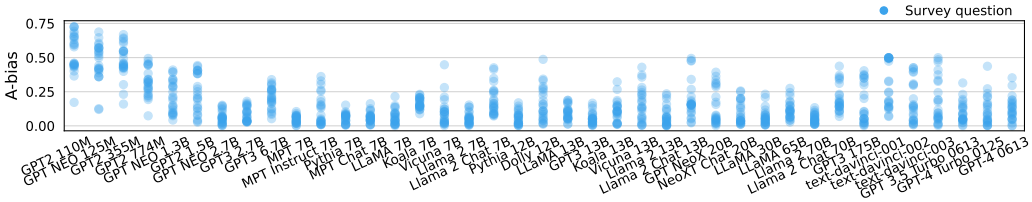


Figure 7: A-bias in language models' survey responses.

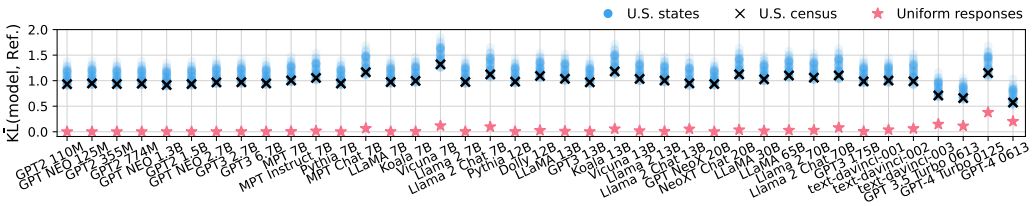


Figure 8: Models' responses are more similar to the uniform baseline than to any of the populations considered.

C.2 A-BIAS OF INSTRUCTION-TUNED MODELS

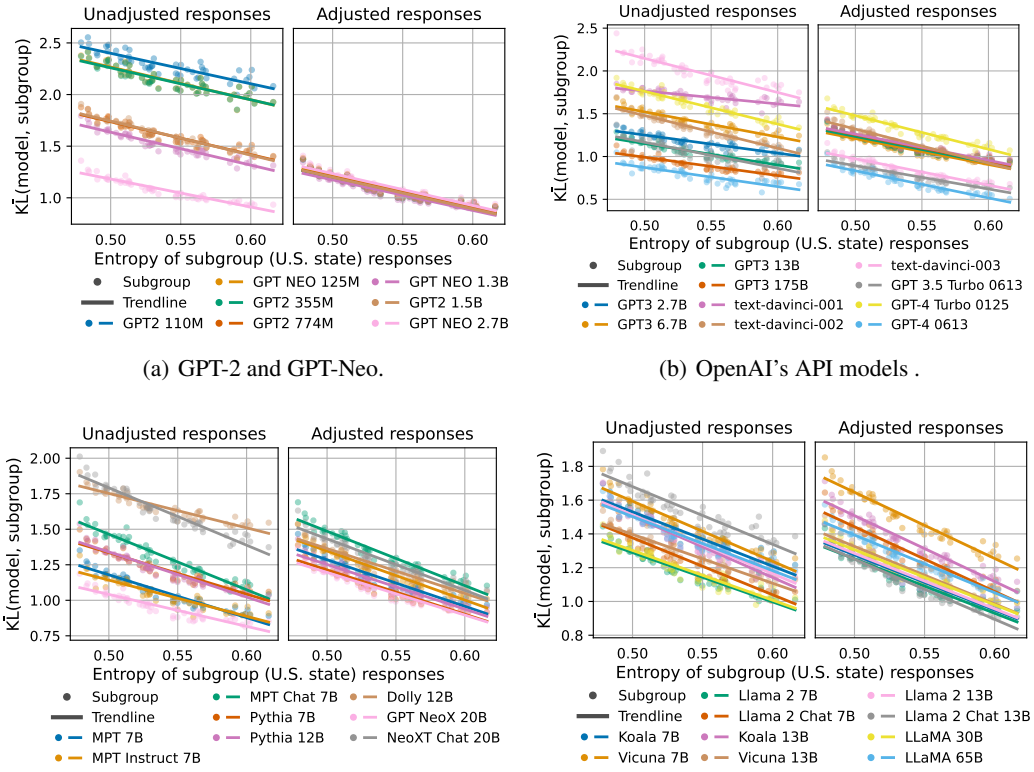
The results in this section complement Section 3.1. We observe that all models exhibit substantial A-bias plotted in Figure 7. This motivates the use of choice-order randomization in order to eliminate confounding due to labeling biases in models' responses.

C.3 ALIGNMENT TO DIFFERENT DEMOGRAPHIC SUBGROUPS

The results in this section complement Section 4.1. After adjusting for choice ordering bias, models' responses are more similar to the uniform baseline than to any of the populations considered, see Figure 8.

C.4 RELATIVE ALIGNMENT ACROSS DEMOGRAPHIC SUBGROUPS

The results presented here complement those of Section 4.1. 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 9(a)), OpenAI's API models (Figure 9(b)), MPT, Pythia, GPT-NeoX and its instruction variants (Figure 9(c)), and LLaMA, Llama 2 and its instruction and chat variants (Figure 9(d)).



(a) GPT-2 and GPT-Neo.

(b) OpenAI's API models .

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

(d) LLaMA, Llama 2 and its instruction variants.

Figure 9: Relative alignment across demographic subgroups for all language models considered.

D 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:

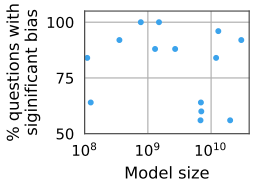
1. We randomize both the order in which choices are presented and the label (i.e., letter) assigned to each answer choice. For example, for the "sex" question, the possible combinations are "A. Male B. Female", "A. Female B. Male", "B. Male A. Female", and "B. Female A. Male". Note that in the experiments 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.

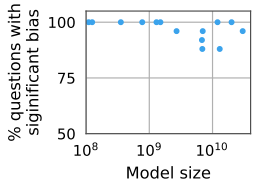
D.1 DISENTANGLING ORDERING BIAS INTO POSITIONING BIAS AND LABELING BIAS

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 (≥ 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.

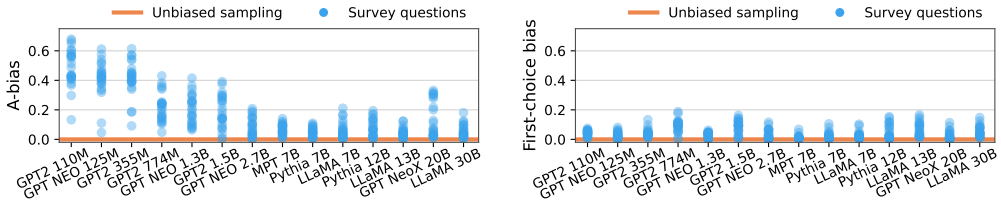


(b) Significant label bias.

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

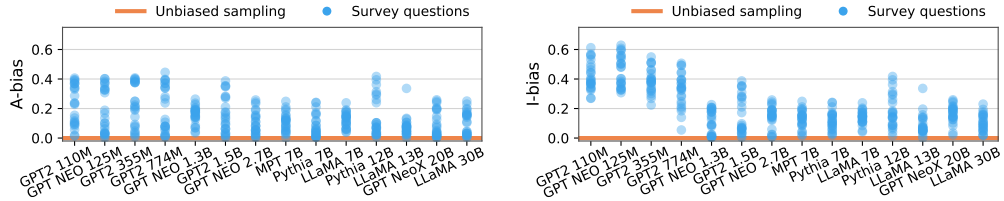
D.2 I-BIAS

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.



(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.



(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.

D.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 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.

(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.

E 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 the repository https://drive.google.com/drive/folders/1HEPo54-G7fthX7JEyws0MuvJFBk8x7Tt?usp=sharing. 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. 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 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\nC. 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}\nAssistant: If had to select one of the options, my answer would be`

Interview-style prompt. We consider the prompt used by Argyle et al. (2022):

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

F RESULTS FOR ATP AND GAS/WVS

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 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 the following repository: `https://drive.google.com/drive/folders/1HEPo54-G7fthX7JEyws0MuvJFBk8x7Tt?usp=sharing`

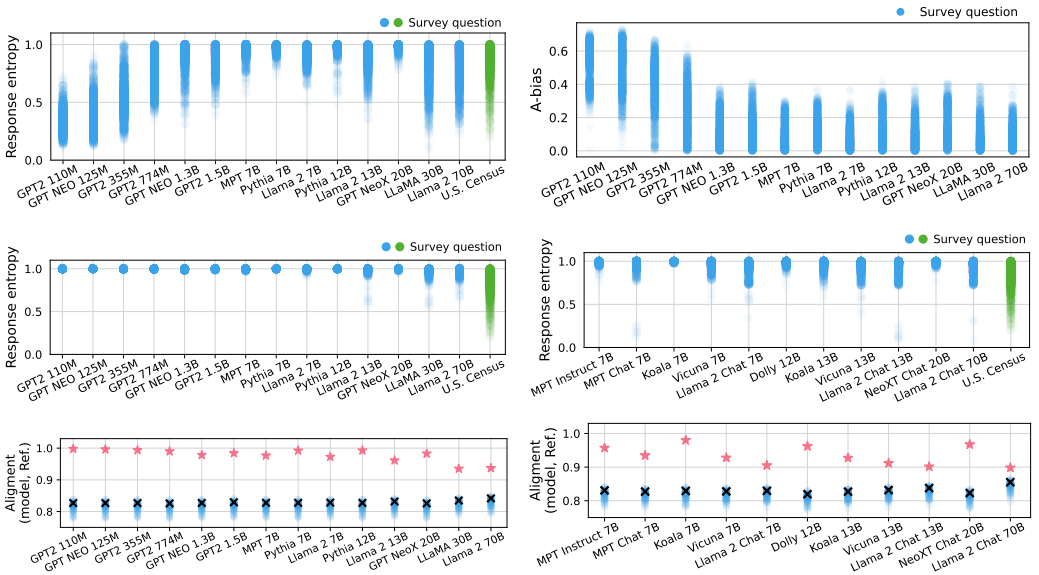


Figure 14: Reproduction of the experiments in Sections 3 and 4 for the ATP surveys.

F.1 ATP SURVEYS

We obtain the ATP survey questions and their corresponding human responses from the OpinionsQA repository.[5] 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.

F.2 GAS AND WVS SURVEYS

We obtain the ATP survey questions and their corresponding human responses from the GlobalOpinionsQA repository.[6] 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.

F.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.

Note that Santurkar et al. (2023) observe that RLHF can result in a "substantial shift [...] towards more liberal, educated, and wealthy [demographic groups]". Our results suggest that this could be an artifact of systematic biases. For the ATP surveys, we observe one outlier for which its alignment *before adjustment* is not correlated with the entropy of subgroup's responses: Llama 2 70B Chat, an RLHF-tuned model. However, after adjustment, Llama 2 70B Chat's alignment trend is remarkably similar to that of Llama 2 70B and all other LLMs, see Figure 17.

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

[6] https://huggingface.co/datasets/Anthropic/llm_global_opinions

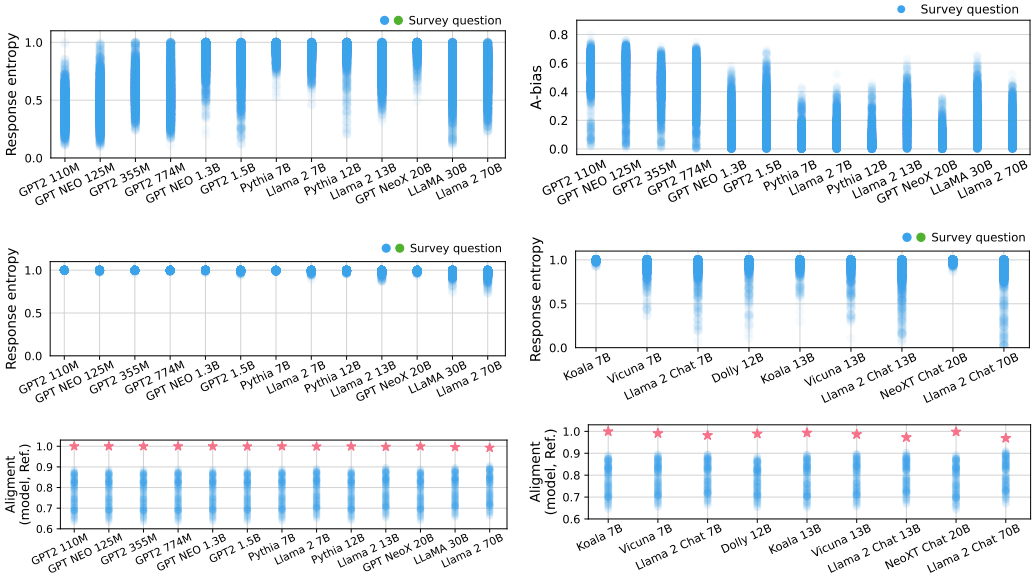
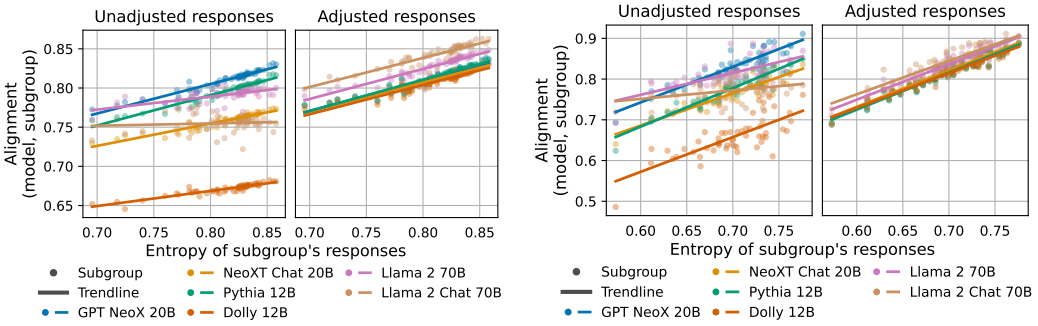


Figure 15: Reproduction of the experiments in Sections 3 and 4 for the GAS/WVS surveys.



(a) ATP surveys.

(b) GAS/WVS surveys.

Figure 16: 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. The alignment between models and a given subpopulation is highly correlated with the entropy of the subpopulations' responses.

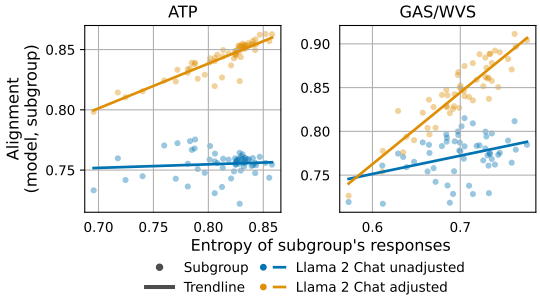


Figure 17: Alignment measures proposed by Santurkar et al. (2023) and Durmus et al. (2023) on ATP and GAS/VVS opinion surveys for Llama 2 70B Chat. The correlation between alignment and the entropy of subgroup's responses is either non-existent or weak before adjustment. However, such correlation is much stronger after adjustment, comparable to that of all other language models, see Figure 16.