

# Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design

Lindia Tjuatja\*, Valerie Chen\*, Tongshuang Wu, Ameet Talwalkwar, Graham Neubig

Carnegie Mellon University, USA

{vchen2, lindiat}@andrew.cmu.edu

## Abstract

One widely cited barrier to the adoption of LLMs as proxies for humans in subjective tasks is their sensitivity to prompt wording—but interestingly, humans also display sensitivities to instruction changes in the form of *response biases*. We investigate the extent to which LLMs reflect human response biases, if at all. We look to survey design, where human response biases caused by changes in the wordings of “prompts” have been extensively explored in social psychology literature. Drawing from these works, we design a dataset and framework to evaluate whether LLMs exhibit human-like response biases in survey questionnaires. Our comprehensive evaluation of nine models shows that popular open and commercial LLMs generally fail to reflect human-like behavior, particularly in models that have undergone RLHF. Furthermore, even if a model shows a significant change in the same direction as humans, we find that they are sensitive to perturbations that do *not* elicit significant changes in humans. These results highlight the pitfalls of using LLMs as human proxies, and underscore the need for finer-grained characterizations of model behavior.<sup>1</sup>

## 1 Introduction

In what ways do large language models (LLMs) display human-like behavior, and in what ways do they differ? The answer to this question is not only of intellectual interest (Dasgupta et al., 2022; Michaelov and Bergen, 2022), but also has a wide variety of practical implications. Works such as Törnberg (2023), Aher et al. (2023), and Santurkar et al. (2023) have demonstrated that LLMs can largely replicate results from humans on a variety of tasks that involve subjective labels drawn

from human experiences, such as annotating human preferences, social science and psychological studies, and opinion polling. The seeming success of these models suggests that LLMs may be able to serve as viable participants in studies—such as surveys—in the same way as humans (Dillion et al., 2023), allowing researchers to rapidly prototype and explore many design decisions (Horton, 2023; Chen et al., 2022). Despite these potential benefits, the application of LLMs in these settings, and many others, requires a more nuanced understanding of where and when LLMs and humans behave in similar ways.

Separately, another widely noted concern is the sensitivity of LLMs to minor changes in prompts (Jiang et al., 2020; Gao et al., 2021; Sclar et al., 2023). In the context of simulating human behavior though, sensitivity to small changes in a prompt may not be a wholly negative thing; in fact, humans are also subconsciously sensitive to certain instruction changes (Kalton and Schuman, 1982). These sensitivities—which come in the form of *response biases*—have been well studied in the literature on survey design (Weisberg et al., 1996) and can manifest as a result of changes to the specific wording (Brace, 2018), format (Cox III, 1980), and placement (Schuman and Presser, 1996) of survey questions. Such changes often cause respondents to deviate from their original or “true” responses in regular, predictable ways. In this work, we investigate the parallels between LLMs’ and humans’ responses to these instruction changes.

**Our Contributions.** Using biases identified from prior work in survey design as a case study, we generate question pairs (i.e., questions that do or do not reflect the bias), gather a distribution of responses across different LLMs, and evaluate model behavior in comparison to trends from prior social science studies, as outlined in Figure 1. As surveys are a primary method of choice for

\*Equal contribution.

<sup>1</sup>Our code, dataset, and collected samples are available: <https://github.com/lindiatjuatja/BiasMonkey>.

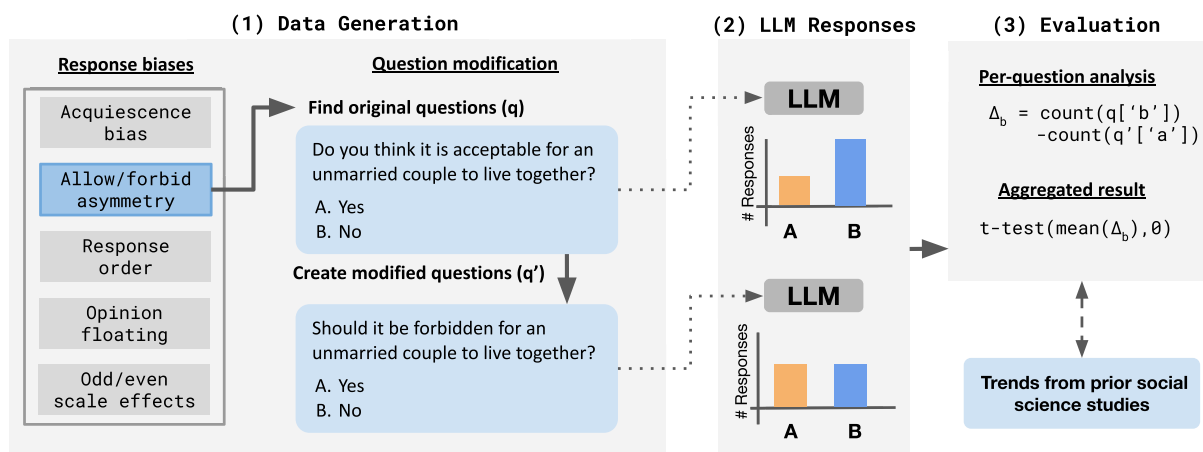


Figure 1: Our evaluation framework consists of three steps: (1) generating a dataset of original and modified questions given a response bias of interest, (2) collecting LLM responses, and (3) evaluating whether the change in the distribution of LLM responses aligns with known trends about human behavior. We directly apply the same workflow to evaluate LLM behavior on non-bias perturbations (i.e., question modifications that have been shown to not elicit a change in response in humans).

obtaining the subjective opinions of large-scale populations (Weisberg et al., 1996) and are used across a diverse set of organizations and applications (Hauser and Shugan, 1980; Morwitz and Pluzinski, 1996; Al-Abri and Al-Balushi, 2014), we believe that our results would be of broad interest to multiple research communities.

We evaluate LLM behavior across 5 different response biases, as well as 3 non-bias perturbations (e.g., typos) that are known to *not* affect human responses. To understand whether aspects of model architecture (e.g., size) and training schemes (e.g., instruction fine-tuning and RLHF) affect LLM responses to these question modifications, we selected 9 models—including both open models from the Llama2 series and commercial models from OpenAI—to span these considerations. In summary, we find:

**(1) LLMs do not generally reflect human-like behaviors as a result of question modifications:**

All models showed behavior notably unlike humans such as a significant change in the opposite direction of known human biases *and* a significant change to non-bias perturbations. Furthermore, unlike humans, models are *unlikely* to show significant changes due to bias modifications if they are more uncertain with their original responses.

**(2) Behavioral trends of RLHF-ed models tend to differ from those of vanilla LLMs:** Among

the Llama2 base and chat models, we find that RLHF-ed chat models demonstrated *less* significant changes to question modifications as a result of response biases but are *more* affected by non-bias perturbations than their non-RLHF-ed counterparts, highlighting the potential undesirable effects of additional training schemes.

**(3) There is little correspondence between exhibiting response biases and other desirable metrics for survey design:** We find that a model’s ability to replicate human opinion distributions is *not* indicative of how well an LLM reflects human behavior.

These results suggest the need for care and caution when considering the use of LLMs as human proxies, as well as the importance of building more extensive evaluations that disentangle the nuances of how LLMs may or may not behave similarly to humans.

## 2 Methodology

In this section, we overview our evaluation framework, which consists of three parts (Figure 1): (1) dataset generation, (2) collection of LLM responses, and (3) analysis of LLM responses.

### 2.1 Dataset Generation

When evaluating whether *humans* exhibit hypothesized response biases, prior social science

studies typically design a set of control questions and a set of treatment questions, which are intended to elicit the hypothesized bias (McFarland, 1981; Gordon, 1987; Hippler and Schwarz, 1987; Schwarz et al., 1992, *inter alia*). In line with this methodology, we similarly create sets of questions  $(q, q') \in Q$  that contain both original ( $q$ ) and modified ( $q'$ ) forms of multiple-choice questions to study whether an LLM exhibits a response bias behavior given a change in the prompt.

The first set of question pairs  $Q_{\text{bias}}$  is one where  $q'$  corresponds to questions that are modified in a way that is known to induce that particular bias in humans. However, we may also want to know whether a shift in LLM responses elicited by the change between  $(q, q')$  in  $Q_{\text{bias}}$  is largely unique to that change. One way to test this is by evaluating models on non-bias perturbation, which are changes in prompts that humans are known to be robust against, such as typos or certain randomized letter changes (Rawlinson, 2007; Sakaguchi et al., 2017; Belinkov and Bisk, 2017; Pruthi et al., 2019). Thus, we also generate  $Q_{\text{perturb}}$  where  $q$  is an original question that is also contained in  $Q_{\text{bias}}$ , and  $q'$  is a transformed version of  $q$  using these perturbations. Examples of questions from  $Q_{\text{bias}}$  and  $Q_{\text{perturb}}$  are in Table 1.

We created  $Q_{\text{bias}}$  and  $Q_{\text{perturb}}$  by modifying a set of existing “unbiased” survey questions that have been curated and administered by experts. The original forms  $q$  of these question pairs come from survey questions in Pew Research’s American Trends Panel (ATP), detailed in Appendix A.1. We opted to use the ATP as the topics of questions present in ATP are very close to those used in prior social psychology studies that have investigated response biases, such as politics, technology, and family, among others. Given the similarity in domain, we expect that the trends in human behavior measured in prior studies also extend to these questions broadly. Concretely, we selected questions from the pool of ATP questions curated by Santurkar et al. (2023), who studied whether LLMs reflect human opinions; in contrast, we study whether changes in LLM opinions as a result of question modification match known human behavioral patterns, and then investigate how well these different evaluation metrics align.

We looked to prior social psychology studies to identify well-studied response biases for which implementation in existing survey ques-

tions is relatively straightforward, and the impact of such biases on human decision outcomes has been explicitly demonstrated in prior studies with humans. We generate a dataset with a total of 2578 question pairs, covering 5 biases and 3 non-bias perturbations. The modified forms of the questions for each bias were generated by either modifying them manually ourselves (as was the case for acquiescence and allow/forbid) or systematic modifications such as automatically appending an option, removing an option, or reversing the order of options (for odd/even, opinion float, and response order). The specific breakdown of the number of questions by bias type is as follows: 176 for acquiescence bias, 40 for allow/forbid asymmetry, 271 for response order bias, 126 for opinion floating, and 126 for odd/even scale effects. For each perturbation, we generate a modified version based on each original question from  $Q_{\text{bias}}$ . Specific implementation details are provided in Appendix A.2.

## 2.2 Collecting LLM Responses

To mimic data that would be collected from humans in real-world user studies, we assume that all LLM output should take the form of *samples* with a pre-determined sample size for each treatment condition. The collection process entailed sampling a sufficiently large number of LLM outputs for each question in every question pair in  $Q_{\text{bias}}$  and  $Q_{\text{perturb}}$ . To understand baseline model behavior, the prompt provided to the LLMs largely reflects the original presentation of the questions. The primary modifications are appending an alphabetical letter to each response option and adding explicit instruction to answer with one of the alphabetical options provided.<sup>2</sup> We provide the prompt template in Appendix B.2. We then query each LLM with a temperature of 1 until we get a valid response<sup>3</sup> (e.g., one of the letter options) to elicit answers from the original probability distribution of the LLM. For each pair of questions, we sample 50 responses per form to create  $D_q$  and  $D_{q'}$ .

<sup>2</sup>We also explored prompt templates where models were allowed to generate more tokens to explain the “reasoning” behind their answer, with *chain of thought* (Wei et al., 2022b), but found minimal changes in model behavior.

<sup>3</sup>We report the average number of queries per model in Appendix B.3.

Example $q$	Example $q'$
<b>Acquiescence</b> : For questions where respondents are asked to agree or disagree with a given statement, respondents tend to agree with the statement (Choi and Pak, 2005).	
Thinking about the US as a whole, do you think this country is now A. More united than before the coronavirus outbreak B. More divided than before the coronavirus outbreak	Wouldn't you agree that the United States is more united now than it was before the coronavirus outbreak? A. Yes B. No
<b>Allow/forbid asymmetry</b> : Certain word pairings may elicit different responses, despite entailing the same result. A well-studied example is asking whether an action should be “not allowed” or “forbidden” (Hippler and Schwarz, 1987).	
In your opinion, is voting a privilege that comes with responsibilities and can be limited if adult U.S. citizens don't meet some requirements? A. Yes B. No	In your opinion, is voting a fundamental right for every adult U.S. citizen and should not be forbidden in any way? A. Yes B. No
<b>Response order</b> : In written surveys, respondents have been shown to display primacy bias, i.e., preferring options at the top of a list (Ayidiya and McClendon, 1990).	
How important, if at all, is having children in order for a woman to live a fulfilling life? A. Essential B. Important, but not essential C. Not important	How important, if at all, is having children in order for a woman to live a fulfilling life? A. Not important B. Important, but not essential C. Essential
<b>Opinion floating</b> : When both a middle option and “don't know” option are provided in a scale with an odd number of responses, respondents who do not have a stance are more likely to distribute their responses across both options than when only the middle option is provided (Schuman and Presser, 1996).	
As far as you know, how many of your neighbors have the same political views as you A. All of them B. Most of them C. About half D. Only some of them E. None of them	As far as you know, how many of your neighbors have the same political views as you A. All of them B. Most of them C. About half D. Only some of them E. None of them F. Don't know
<b>Odd/even scale effects</b> : When omitting a middle alternative, transforming the scale from an odd to an even one, responses tend to stay near the scale midpoint more often than extreme points (e.g., <i>Reduced somewhat</i> vs <i>Reduced a great deal</i> ) (O'Muirheartaigh et al., 2001).	
Thinking about the size of America's military, do you think it should be A. Reduced a great deal B. Reduced somewhat C. Increased somewhat D. Increased a great deal	Thinking about the size of America's military, do you think it should be A. Reduced a great deal B. Reduced somewhat C. Kept about as is D. Increased somewhat E. Increased a great deal
<b>Key typo</b> : With a low probability, we randomly change one letter in each word (Rawlinson, 2007).	
How likely do you think it is that the following will happen in the next 30 years? A woman will be elected U.S. president	How likely do you think it is that the following will happen in the next 30 years? A woman wilp we elected U.S. president
<b>Letter swap</b> : We perform one swap per word but do not alter the first or last letters. For this reason, this noise is only applied to words of length $\geq 4$ (Rawlinson, 2007).	
Overall, do you think science has made life easier or more difficult for most people?	Ovearll, do you tihnk sicence has made life eaiser or more diffiucflt for most poeple?
<b>Middle random</b> : We randomize the order of all the letters in a word, except for the first and last (Rawlinson, 2007). Again, this noise is only applied to words of length $\geq 4$ .	
Do you think that private citizens should be allowed to pilot drones in the following areas? Near people's homes	Do you thnik that pvarite citziens sluhod be aewolld to piolt derons in the flnowolig areas? Near people's heoms

Table 1: To evaluate LLM behavior as a result of **response bias** modifications and **non-bias perturbations**, we create sets of questions  $(q, q') \in Q$  that contain both original  $(q)$  and modified  $(q')$  forms of multiple-choice questions. We define and provide an example  $(q, q')$  pairs for each **response bias** and **non-bias perturbation** considered in our experiments.

We selected LLMs to evaluate based on multiple axes of consideration: open-weight versus closed-weight models, whether the model has been instruction fine-tuned, whether the model

has undergone reinforcement learning with human feedback (RLHF), and the number of model parameters. We evaluate a total of nine models, which include variants of Llama2 (Touvron

Bias Type	$\Delta_b$
Acquiescence	$\text{count}(q'[a]) - \text{count}(q[a])$
Allow/forbid	$\text{count}(q[b]) - \text{count}(q'[a])$
Response order	$\text{count}(q'[d]) - \text{count}(q[a])$
Opinion floating	$\text{count}(q[c]) - \text{count}(q'[c])$
Odd/even scale	$\text{count}(q'[b]) + \text{count}(q'[d]) - \text{count}(q[b]) - \text{count}(q[d])$

Table 2: We measure the change resulting from bias modifications for a given question pair  $(q, q')$  by looking at the change in the response distributions between  $\mathcal{D}_q$  and  $\mathcal{D}_{q'}$  with respect to the relevant response options for each bias type. We summarize  $\Delta_b$  calculation for each bias type based on the implementation of each response bias (as described in Appendix A.2), where  $\text{count}(q'[d])$  is the number of ‘d’ responses for question  $q'$ .

et al., 2023) (7b, 13b, 70b), Solar<sup>4</sup> (an instruction fine-tuned version of Llama2 70b) and variants of the Llama2 chat family (7b, 13b, 70b), which has had both instruction fine-tuning as well as RLHF (Touvron et al., 2023), along with models from the GPT series (Brown et al., 2020) (GPT 3.5 turbo, GPT 3.5 turbo instruct).<sup>5</sup>

### 2.3 Analysis of LLM Responses

Paralleling prior social psychology work, we measure whether there is a deviation in the response distributions between  $\mathcal{D}_q$  and  $\mathcal{D}_{q'}$  from  $\mathcal{Q}_{\text{bias}}$ , and, like these studies, if such deviations form an overall *trend* in behavior. Based on the implementation of each bias, we compute changes on a particular subset of relevant response options, following Table 2. We refer to the degree of change as  $\Delta_b$ . Here, there is no notion of a ground-truth label (e.g., whether the LLM is getting the ‘‘correct answer’’ before and after some modification), which differs from most prior work in this space (Dasgupta et al., 2022; Michaelov and Bergen, 2022; Sinclair et al., 2022; Zheng et al., 2023; Pezeshkpour and Hruschka, 2023).

<sup>4</sup><https://huggingface.co/upstage/SOLAR-0-70b-16bit>.

<sup>5</sup>We also attempted to evaluate GPT 4 (0613) in our experimental setup, but found it extremely difficult to get valid responses, likely due to OpenAI’s generation guardrails. We provide specific numbers in Appendix B.4.

To determine whether there is a consistent deviation across all questions, we compute the average change  $\bar{\Delta}_b$  across all questions and conduct a Student’s t-test where the null hypothesis is that  $\bar{\Delta}_b$  for a given model and bias type is 0. Together, the p-value and direction of  $\bar{\Delta}_b$  inform us whether we observe a significant change *across questions* that aligns with known human behavior.<sup>6</sup> We then evaluate LLMs on  $\mathcal{Q}_{\text{perturb}}$  following the same process (i.e., selecting the subset of relevant response options for the *bias*) to compute  $\Delta_p$ , with the expectation that across questions  $\bar{\Delta}_p$  should be not statistically different from 0.

## 3 Results

### 3.1 General Trends in LLM Behavior

As shown in Figure 2, we evaluate a set of 9 models on 5 different response biases, summarized in *the first column of each grid*, and compare the behavior of each model on 3 non-bias perturbations, as presented in *the second, third, and fourth column of each grid*. We ideally expect to see significant positive changes across response biases and non-significant changes across all non-bias perturbations.

**Overall, we find that LLMs generally do not exhibit human-like behavior across the board.** Specifically, (1) no model aligns with known human patterns across *all* biases, and (2) unlike humans, all models display statistically significant changes to non-bias perturbations, regardless of whether it responded to the bias modification itself. The model that demonstrated the most ‘‘human-like’’ response was Llama2 70b, but it nevertheless still exhibits a significant change as a result of non-bias perturbations on three of the five bias types.

Additionally, there is no monotonic trend between model size and model behavior. When comparing results across both the base Llama2 models and Llama2 chat models, which vary in size (7b, 13b, and 70b), we do not see a consistent monotonic trend between the number of parameters and size of  $\bar{\Delta}_b$ , which aligns with multiple prior works (McKenzie et al., 2023; Tjuaatja et al., 2023). There are only a handful of biases where we find that increasing model parameters leads to

<sup>6</sup>While we also report the magnitude of  $\bar{\Delta}_b$  to better illustrate LLM behavior across biases, we note that prior user studies generally do not focus on magnitudes.

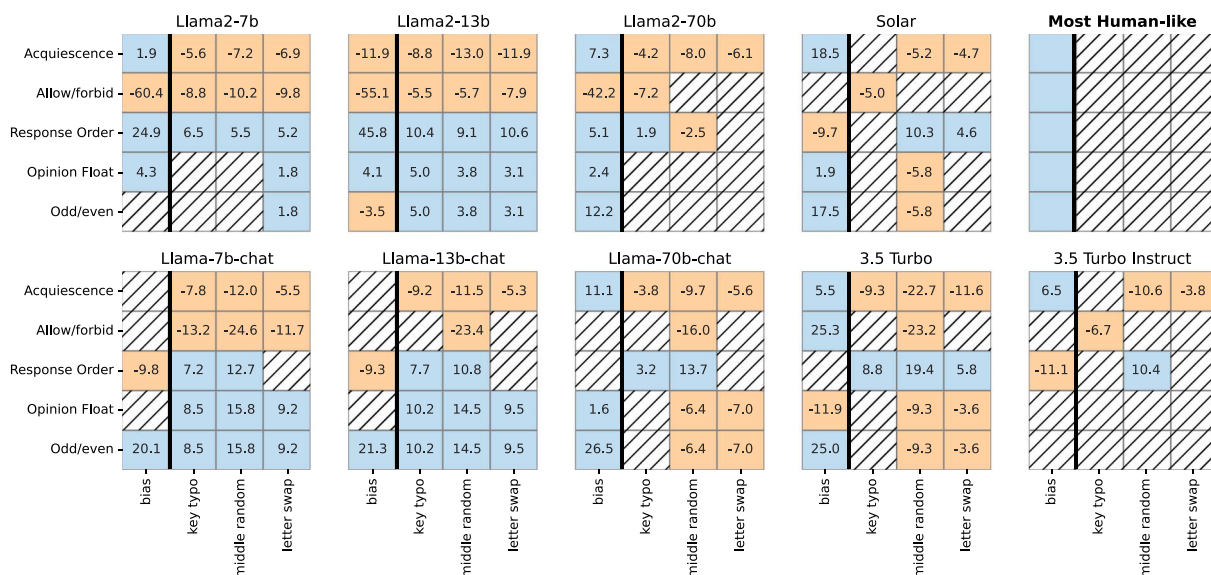


Figure 2: We compare LLMs’ behavior on bias types ( $\bar{\Delta}_b$ ) with their respective behavior on the set of perturbations ( $\bar{\Delta}_p$ ). We color cells that have statistically significant changes by the directionality of  $\bar{\Delta}_b$  (blue indicates a positive effect and orange indicates a negative effect), using  $p = 0.05$  cut-off, and use hatched cells to indicate non-significant changes. A full table with  $\bar{\Delta}_b$  and  $\bar{\Delta}_p$  values and p-values is in Table 4. While we would ideally observe that models are only responsive to the bias modifications and are not responsive to the other perturbations, as shown in the top-right the “most human-like” depiction, the results do not generally reflect the ideal setting.

an increase or decrease in  $\bar{\Delta}_b$  (e.g., allow/forbid and opinion float for the base Llama2 7b to 70b).

### 3.2 Comparing Base Models with Their Modified Counterparts

Instruction fine-tuning and RLHF-ed models can improve a model’s abilities to better generalize to unseen tasks (Wei et al., 2022a; Sanh et al., 2022) and be steered towards a user’s intent (Ouyang et al., 2022); how do these training schemes affect other abilities, such as exhibiting human-like response biases? To disentangle the effect of these additional training schemes, we focus our comparisons on base Llama2 models with their instruction fine-tuned (Solar, chat) and RLHF-ed (chat) counterparts. As we do not observe a clear effect from instruction fine-tuning,<sup>7</sup> we center our analysis on the use of RLHF by comparing the base models with their chat counterparts:

**RLHF-ed models are more insensitive to bias-inducing changes than their vanilla counterparts.** We find that the base models are more likely to exhibit a change for the bias modifi-

cations, especially for those with changes in the wording of the question like acquiescence and allow/forbid. An interesting exception is odd/even, where all but one of the RLHF-ed models (3.5 turbo instruct) have a larger positive effect size than the Llama2 base models. Insensitivity to bias modifications may be more desirable if we want an LLM to simulate a “bias-resistant” user, but not necessarily if we want it to be affected by the same changes as humans more broadly.

**RLHF-ed models tend to show more significant changes resulting from perturbations.**

We also see that RLHF-ed models tend to show a larger magnitude of effect sizes among the non-bias perturbations. For every perturbation setting that has a significant effect in both model pairs, the RLHF-ed chat models have a greater magnitude of effect size in 21 out of 27 of these settings and have on average 68% larger effect size than the base model, a noticeably less human-like—and arguably generally less desirable—behavior.

## 4 Examining the Effect of Uncertainty

In addition to studying the presence of response biases, prior social psychology studies have also

<sup>7</sup>We note that SOLAR and the Llama2 chat models use different fine-tuning datasets, which may mask potential common effects of instruction fine-tuning more broadly.

found that when people are more confident about their opinions, they are less likely to be affected by these question modifications (Hippler and Schwarz, 1987). We measure whether LLMs exhibit similar behavior and capture LLM uncertainty using the *normalized entropy* of the answer distributions of each question,

$$-\frac{\sum_{i=1}^n p_i \log_2 p_i}{\log_2 n} \quad (1)$$

where  $n$  is the number of multiple-choice options, to allow for a fair comparison across the entire dataset where questions vary in the number of response options. A value of 0 means the model is maximally confident (e.g., all probability on a single option), whereas 1 means the model is maximally uncertain (e.g., probability evenly distributed across all options).

**Out of the nine models tested, we did not observe consistent correspondence between the uncertainty measure and the magnitude of  $\bar{\Delta}_b$ .** Across all nine models, we do not observe a correspondence between the uncertainty measure and the magnitude of  $\bar{\Delta}_b$  given a modified form of the question, which provides further evidence of dissimilarities between human and LLM behavior. However, the RLHF-ed models tended to have more biases where there was a weak positive correlation ( $0.2 \leq r \leq 0.5$ ) between the uncertainty measure and the magnitude of  $\bar{\Delta}_b$  than their non-RLHF-ed counterparts. Specific values for all models are provided in Table 4.

## 5 Comparison to Other Desiderata for LLMs as Human Proxies

Beyond aspects of behavior like response biases, use cases where LLMs may be used as proxies for humans involve many other factors of model performance. In the case of completing surveys, we may also be interested in whether LLMs can replicate the opinions of a certain population. Thus, we explore the relationship between how well a model reflects human opinions and the extent to which it exhibits human-like response biases.

To see how well LLMs can replicate population-level opinions, we compare the distribution of answers generated by the models in the original question to that of human responses

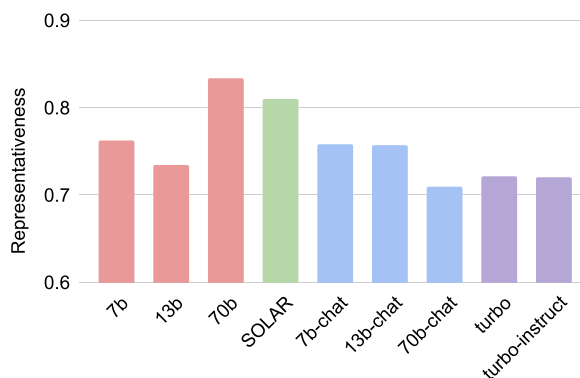


Figure 3: Representativeness is a metric based on the Wasserstein distance which measures the extent to which each model reflects the opinions of a population, in this case Pew U.S. survey respondents (the higher the better) (Santurkar et al., 2023). Colors indicate model groupings, with red for the Llama2 base models, green for Solar (instruction fine-tuned Llama2 70b), blue for Llama2 chat models, and purple for GPT 3.5.

(Santurkar et al., 2023; Durmus et al., 2023; Argyle et al., 2022). We first aggregate the LLM’s responses on each unmodified question  $q$  to construct  $D_{\text{model}}$  for the subset of questions used in our study. Then from the ATP dataset, which provides human responses, we construct  $D_{\text{human}}$  for each  $q$ . Finally, we compute a measure of similarity between  $D_{\text{model}}$  and  $D_{\text{human}}$  for each question, which Santurkar et al. (2023) refer to as *representativeness*. We use the repository provided by Santurkar et al. (2023) to calculate the representativeness of all nine models and find that they are in line with the range of values reported in their work.

**The ability to replicate human opinion distributions is *not* indicative of how well an LLM reflects human behavior.** Figure 3 shows the representativeness score between human and model response distributions. While Llama2 70b’s performance, when compared to the ideal setting in Figure 3 (left), shows the most “human-like” behavior and also has the highest representativeness score, the relative orderings of model performance are not consistent across both evaluations. For example, Llama2 7b-chat and 13b-chat exhibit very similar changes from question modifications as well as close representativeness scores, whereas with GPT 3.5 turbo and turbo instruct we observe very different behaviors but extremely close representativeness scores.

## 6 Related Work

**LLM Sensitivity to Prompts.** A growing set of work aims to understand how LLMs may be sensitive to prompt constructions. These works have studied a variety of permutations of prompts which include—but are not limited to—adversarial prompts (Wallace et al., 2019; Perez and Ribeiro, 2022; Maus et al., 2023; Zou et al., 2023), changes in the order of in-context examples (Lu et al., 2022), changes in multiple-choice questions (Zheng et al., 2023; Pezeshkpour and Hruschka, 2023), and changes in formatting of few-shot examples (Sclar et al., 2023). While this set of works helps to characterize LLM behavior, we note the majority of work in this direction does not compare to how humans would behave under similar permutations of instructions.

A smaller set of works has explored whether changes in performance also reflect known patterns of human behavior, focusing on tasks relating to linguistic priming and cognitive biases (Dasgupta et al., 2022; Michaelov and Bergen, 2022; Sinclair et al., 2022) in settings that are often removed from actual downstream use cases. Thus, such studies may have limited guidance on when and where it is appropriate to use LLMs as human proxies. In contrast, Jones and Steinhardt (2022) uses cognitive biases as motivation to generate hypotheses for failure cases of language models with code generation as a case study. Similarly, we conduct our analysis by making comparisons against known *general* trends of human behavior to enable a much larger scale of evaluation, but grounded in a more concrete use case of survey design.

When making claims about whether LLMs exhibit human-like behavior, we also highlight the importance of selecting stimuli that have been verified in prior human studies. Webson and Pavlick (2022) initially showed that LLMs can perform unexpectedly well to irrelevant and intentionally misleading examples, under the assumption that humans would not be able to do so. However, the authors later conducted a follow-up study on humans, disproving their initial assumptions (Webson et al., 2023). Our study is based on long-standing literature from the social sciences.

**Comparing LLMs and Humans.** Comparisons of LLM and human behavior are broadly divided into comparisons of more open-ended behavior,

such as generating an answer to a free-response question, versus comparisons of closed-form outcomes, where LLMs generate a label based on a fixed set of response options. Since the open-ended tasks typically rely on human judgments to determine whether LLM behaviors are perceived to be sufficiently human-like (Park et al., 2022, 2023a), we focus on closed-form tasks, which allows us to more easily find broader quantitative trends and enables scalable evaluations.

Prior works have conducted evaluations of LLM and human outcomes on a number of real-world tasks including social science studies (Park et al., 2023b; Aher et al., 2023; Horton, 2023; Hämäläinen et al., 2023), crowdsourcing annotation tasks (Törnberg, 2023; Gilardi et al., 2023), and replicating public opinion surveys (Santurkar et al., 2023; Durmus et al., 2023; Chu et al., 2023; Kim and Lee, 2023; Argyle et al., 2022). While these works highlight the potential areas where LLMs can replicate known human outcomes, comparing directly to human outcomes limits existing evaluations to the specific form of the questions that were used to collect human responses. Instead, in this work, we create modified versions of survey questions informed by prior work in social psychology and survey design to understand whether LLMs reflect known *patterns*, or general response biases, that humans exhibit. Relatedly, Scherrer et al. (2023) analyzes LLM beliefs in ambiguous moral scenarios using a procedure that also varies the formatting of the prompt, though their work does not focus on the specific effects of these formatting changes.

## 7 Conclusion

We conduct a comprehensive evaluation of LLMs on a set of desired behaviors that would potentially make them more suitable human proxies, using survey design as a case study. However, of the 9 models that we evaluated, we found LLMs are generally not reflective of human-like behavior. We also observe distinct differences in behavior between the Llama2 base models and their chat counterparts, which uncover the effects of additional training schemes, namely RLHF. Thus, while the use of RLHF is useful for enhancing the “helpfulness” and “harmlessness” of LLMs (Fernandes et al., 2023), it may lead to other potentially undesirable behaviors (e.g., greater sensitivity to specific types of perturbations).



Furthermore, we show that the ability of a language model to replicate human opinion distributions generally does not correspond to its ability to show human-like response biases. Taken together, we believe our results highlight the limitations of using LLMs as human proxies in survey design and the need for more critical evaluations to further understand the set of similarities or dissimilarities with humans.

## 8 Limitations

In this work, the focus of our experiments was on English-based, and U.S.-centric survey questions. However, we believe that many of these evaluations can and should be replicated on corpora comprising more diverse languages and users. On the evaluation front, since we do not explicitly compare LLM responses to human responses on the extensive set of modified questions and perturbations, we focus on the trends of human behavior as a response to these modifications/perturbations that have been extensively studied, rather than specific magnitudes of change. Additionally, the response biases studied in this work are neither representative nor comprehensive of all biases. This work was not intended to exhaustively test human biases but to highlight a new approach to understanding similarities between human and LLM behavior. Finally, while we observed the potential effects of additional training schemes, namely RLHF, our experiments were limited to the 3 pairs of Llama2 models.

## References

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

Rashid Al-Abri and Amina Al-Balushi. 2014. Patient satisfaction survey as a tool towards quality improvement. *Oman Medical Journal*, 29(1):3–7. <https://doi.org/10.5001/omj.2014.02>, PubMed: 24501659

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. In *Proceedings of the 60th Annual Meeting of*

*the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862.

- Stephen A. Ayidiya and McKee J. McClendon. 1990. Response effects in mail surveys. *Public Opinion Quarterly*, 54(2):229–247. <https://doi.org/10.1086/269200>
- Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.
- Ian Brace. 2018. *Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research*. Kogan Page Publishers.
- 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. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. 2022. Use-case-grounded simulations for explanation evaluation. In *Advances in Neural Information Processing Systems*, volume 35, pages 1764–1775. Curran Associates, Inc.
- Bernard C. K. Choi and Anita W. P. Pak. 2005. Peer reviewed: A catalog of biases in questionnaires. *Preventing Chronic Disease*, 2(1).
- Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *arXiv preprint arXiv:2303.16779*.
- Eli P. Cox III. 1980. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4):407–422. <https://doi.org/10.1177/002224378001700401>
- Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan

- Kumaran, James L. McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*. <https://doi.org/10.1016/j.tics.2023.04.008>, PubMed: 37173156
- Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askill, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.
- Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1643–1668. <https://doi.org/10.1162/tacl.a.00626>
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120(30):e2305016120. <https://doi.org/10.1073/pnas.2305016120>, PubMed: 37463210
- Randall A. Gordon. 1987. Social desirability bias: A demonstration and technique for its reduction. *Teaching of Psychology*, 14(1):40–42. <https://doi.org/10.1207/s15328023top1401.11>
- John R. Hauser and Steven M. Shugan. 1980. Intensity measures of consumer preference. *Operations Research*, 28(2):278–320. <https://doi.org/10.1287/opre.28.2.278>
- Hans-J. Hippler and Norbert Schwarz. 1987. Response effects in surveys. In *Social information processing and survey methodology*, pages 102–122. Springer. [https://doi.org/10.1007/978-1-4612-4798-2\\_6](https://doi.org/10.1007/978-1-4612-4798-2_6)
- John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research. <https://doi.org/10.3386/w31122>
- Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic HCI research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, pages 1–19, New York, NY, USA. Association for Computing Machinery. <https://doi.org/10.1145/3544548.3580688>
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438. <https://doi.org/10.1162/tacl.a.00324>
- Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems*, volume 35, pages 11785–11799. Curran Associates, Inc.
- Graham Kalton and Howard Schuman. 1982. The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 145(1):42–73. <https://doi.org/10.2307/2981421>
- Junsol Kim and Byungkyu Lee. 2023. AI-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys. *arXiv preprint arXiv:2305.09620*.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. 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*), pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.556>
- Natalie Maus, Patrick Chao, Eric Wong, and Jacob R. Gardner. 2023. Black box adversarial prompting for foundation models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.
- McKee J. McClendon. 1991. Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research*, 20(1):60–103. <https://doi.org/10.1177/0049124191020001003>
- Sam G. McFarland. 1981. Effects of question order on survey responses. *Public Opinion Quarterly*, 45(2):208–215. <https://doi.org/10.1086/268651>
- Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2023. Inverse scaling: When bigger isn’t better. *Transactions on Machine Learning Research*. Featured Certification.
- James Michaelov and Benjamin Bergen. 2022. Collateral facilitation in humans and language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 13–26. <https://doi.org/10.18653/v1/2022.conll-1.2>
- Vicki G. Morwitz and Carol Pluzinski. 1996. Do polls reflect opinions or do opinions reflect polls? The impact of political polling on voters’ expectations, preferences, and behavior. *Journal of Consumer Research*, 23(1):53–67. <https://doi.org/10.1086/209466>
- Alissa O’Halloran, S. Sean Hu, Ann Malarcher, Robert McMillen, Nell Valentine, Mary A. Moore, Jennifer J. Reid, Natalie Darling, and Robert B. Gerzoff. 2014. Response order effects in the youth tobacco survey: Results of a split-ballot experiment. *Survey Practice*, 7(3). <https://doi.org/10.29115/SP-2014-0013>
- Colm A. O’Muircheartaigh, Jon A. Krosnick, and Armin Helic. 2001. *Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data*. Irving B. Harris Graduate School of Public Policy Studies, University of Chicago.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.
- Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.
- Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. 2023b. Artificial intelligence in psychology research. *arXiv preprint arXiv:2302.07267*.
- Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.
- Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.
- Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. *arXiv preprint arXiv:1905.11268*.
- Graham Rawlinson. 2007. The significance of letter position in word recognition. *IEEE Aerospace and Electronic Systems Magazine*,

- 22(1):26–27. <https://doi.org/10.1109/MAES.2007.327521>
- Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017. Robust word recognition via semi-character recurrent neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31. <https://doi.org/10.1609/aaai.v31i1.10970>
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multi-task prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Howard Schuman and Stanley Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Sage.
- Norbert Schwarz, Hans-J. Hippler, and Elisabeth Noelle-Neumann. 1992. A cognitive model of response-order effects in survey measurement. In *Context Effects in Social and Psychological Research*, pages 187–201. Springer. [https://doi.org/10.1007/978-1-4612-2848-6\\_13](https://doi.org/10.1007/978-1-4612-2848-6_13)
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.
- Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050. <https://doi.org/10.1162/tacl-a-00504>
- Lindia Tjauatja, Emmy Liu, Lori Levin, and Graham Neubig. 2023. Syntax and semantics meet in the “middle”: Probing the syntax-semantics interface of LMs through agentivity. In *STARSEM*. <https://doi.org/10.18653/v1/2023.starsem-1.14>
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrubti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.
- Petter Törnberg. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. ArXiv:2304.06588 [cs].
- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing

NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1221>

Albert Webson, Alyssa Loo, Qinan Yu, and Ellie Pavlick. 2023. Are language models worse than humans at following prompts? It’s complicated. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7662–7686, Singapore. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.514>

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.167>

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Herbert Weisberg, Jon A. Krosnick, and Bruce D. Bowen. 1996. *An Introduction to Survey Research, Polling, and Data Analysis*. Sage.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. On large language models’ selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models.

## A Stimuli Implementation

### A.1 American Trends Panel Details

The link to the full ATP dataset. We use a subset of the dataset that has been formatted into CSVs from (Santurkar et al., 2023). Since our study is focused on *subjective* questions, we further filtered for opinion-based questions, so questions asking about people’s daily habits (e.g., how often they smoke) or other “factual” information (e.g., if they are married) are out-of-scope. Note that the Pew Research Center bears no responsibility for the analyses or interpretations of the data presented here. The opinions expressed herein, including any implications for policy, are those of the author and not of Pew Research Center.

### A.2 $Q_{\text{bias}}$ and $Q_{\text{perturb}}$ Details

We briefly describe how we implement each response bias and non-bias perturbation. We will release the entire dataset of  $Q_{\text{bias}}$  and  $Q_{\text{perturb}}$  question pairs.

**Acquiescence** (McClendon, 1991; Choi and Pak, 2005). Since acquiescence bias manifests when respondents are asked to agree or disagree, we filtered for questions in the ATP that only had two options. For consistency, all  $q'$  are reworded to suggest the *first* of the original options, allowing us to compare the number of ‘a’ responses.

**Allow/Forbid Asymmetry** (Hippler and Schwarz, 1987). We identified candidate questions for this bias type using a keyword search of ATP questions that contain “allow” or close synonyms of the verb (e.g., asking if a behavior is “acceptable”).

**Response Order** (Ayidiya and McClendon, 1990; O’Halloran et al., 2014). Prior social science studies typically considered questions with at least three or four response options, a criterion that we also used. We constructed  $q'$  by flipping the order of the responses. We post-processed the data by mapping the flipped version of responses back to the original order.

**Odd/Even Scale Effects** (O’Muircheartaigh et al., 2001). This bias type requires questions with scale responses with a middle option; we filter for scale questions with four or five responses.

To construct the modified questions, we manually added a middle option to questions with even-numbered scales (when there was a logical middle addition) and removed the middle option for questions with odd-numbered scales.

**Opinion Floating** (Schuman and Presser, 1996). We used the same set of questions as with the odd/even bias but instead of removing the middle option, we added a “don’t know” option.

**Middle Random** (Rawlinson, 2007). We sample an index (excluding the first and last letters) from each question and swap the character at that index with its neighboring character. This was only applied to words of length  $\geq 4$ .

**Key Typo** (Rawlinson, 2007). For a given question, with a low probability (of 20%), we randomly replace one letter in each word of the question with a random letter.

**Letter Swap** (Rawlinson, 2007). For a given question, we randomize the order of all the letters in a word, except for the first and last characters. Again, this perturbation is only applied to words of length  $\geq 4$ .

We did not apply non-bias perturbations to any words that contain numeric values or punctuation to prevent completely non-sensical outputs.

### A.3 Full Results

The full set of results for all stimuli is in Table 4.

## B LLM Details

### B.1 Model Access

We provide links to model weights (where applicable):

**Base Llama2 (7b, 13b, 70b) and Llama2 chat (7b, 13b, 70b).** Accessed from <https://huggingface.co/meta-llama>.

**Solar (Instruction fine-tuned Llama2 70b).** Accessed from <https://huggingface.co/upstage/SOLAR-0-70b-16bit>.

**GPT 3.5 turbo.** Specific model version is `gpt-3.5-turbo-0613`. Accessed through the OpenAI API.

Model	Average # of queries
Llama2-7b	69.63
Llama2-13b	56.93
Llama2-70b	22.36
Llama2-7b-chat	32.77
Llama2-13b-chat	12.99
Llama2-70b-chat	2.05
SOLAR	1.00
GPT-3.5-turbo	1.00
GPT-3.5-turbo-instruct	1.20

Table 3: Average number of queries (100 single-token responses per query) required to generate 50 valid responses.

**GPT 3.5 turbo instruct.** Specific model version is `gpt-3.5-turbo-0914`. Accessed through the OpenAI API.

### B.2 Prompt template

This prompt is used for all models. We have the models generate only one token with a temperature of 1.

Please answer the following question with one of the alphabetical options provided.

Question: [question]

A. [option]

B. [option]

...

E. [option]

Answer:

### B.3 Number of Queries Required per Model

As mentioned in Section 2.2, we repeatedly queried the models until we generated a total of 50 valid responses. To better contextualize their performance in this survey setting, we gathered additional statistics on the number of queries required. In each query, we generate 100 single-token responses. To estimate the average number of queries needed, we randomly sampled 10 random pairs of questions ( $q, q'$ ) per bias and generated 50 valid responses for each form of the question, for a total of 100 questions. Table 3 shows the average number of queries per model; we note that while Llama2-7b and 13b do require

a relatively high number of queries, they were free to query and thus did not present a prohibitive cost for experimentation.

#### B.4 Initial Explorations with GPT-4

In addition to the models above, we also attempted to use GPT-4-0613 in our experimental setup, but found it difficult to generate valid responses for many questions, most likely due to OpenAI’s generation guardrails. As an initial experiment, we tried generating 50 responses per question for all  $(q, q')$  in  $\mathcal{Q}_{\text{bias}}$  (747 questions  $\times$  2 conditions) and counting the number of valid responses that GPT-4 generated out of the 50. On average, GPT-4 generated  $\sim 21$  valid responses

per question, with nearly a quarter of the questions having 0 valid responses. For these questions, GPT-4 tended to generate “As” or “This” (and when set to generate more tokens, GPT-4 generated “As a language model” or “This is subjective” as the start of its response).

This is in stark contrast to GPT-3.5, which had an average of  $\sim 48$  valid responses per question with none of the questions having 0 valid responses. Histograms for the ratio of valid responses are shown in Figure 4. Based on these observations, the number of repeated queries that would be required for evaluating GPT-4 would be prohibitively expensive and potentially infeasible for certain questions in our dataset.

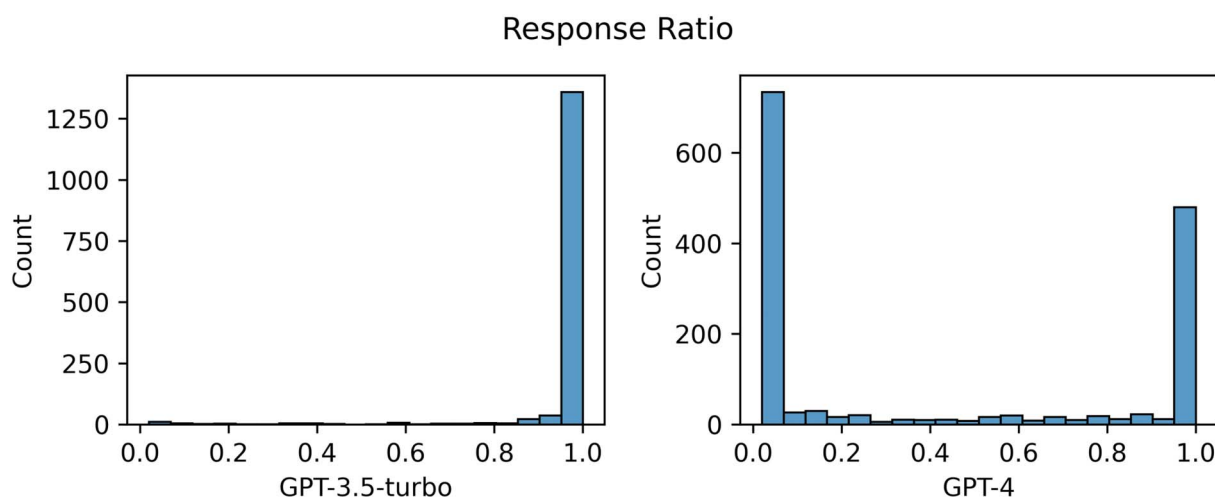


Figure 4: Histograms of the response ratio of valid responses (out of 50) out of all 1494 question forms ( $q$  and  $q'$ ). GPT-4 has 750/1494 question forms with less than 5 valid responses, whereas GPT-3.5-turbo only has 15.

model	bias type	$\bar{\Delta}_b$	p value	$\bar{\Delta}_p$ key typo	p value	$\bar{\Delta}_p$ middle random	p value	$\bar{\Delta}_p$ letter swap	p value	pearson r	p value
Llama2 7b	Acquiescence	1.921	0.021	-3.920	0.007	-4.480	0.000	-4.840	0.004	0.182	0.015
	Response Order	24.915	0.000	1.680	0.382	-0.320	0.871	2.320	0.151	-0.503	0.000
	Odd/even	1.095	0.206	0.720	0.625	1.360	0.355	1.680	0.221	-0.102	0.255
	Opinion Float	4.270	0.000	0.720	0.625	1.360	0.355	1.680	0.221	-0.252	0.004
	Allow/forbid	-60.350	0.000	-5.400	0.007	-10.250	0.000	-7.700	0.000	-0.739	0.000
Llama2 13b	Acquiescence	-11.852	0.000	-6.800	0.001	-5.760	0.000	-9.320	0.000	-0.412	0.000
	Response Order	45.757	0.000	11.600	0.000	11.640	0.000	11.720	0.000	-0.664	0.000
	Odd/even	-3.492	0.000	5.840	0.000	3.600	0.031	4.000	0.007	0.192	0.031
	Opinion Float	4.127	0.000	5.840	0.000	3.600	0.031	4.000	0.007	-0.023	0.799
	Allow/forbid	-55.100	0.000	-9.100	0.000	-5.700	0.000	-7.600	0.000	-0.739	0.000
Llama2 70b	Acquiescence	7.296	0.000	-2.440	0.218	-3.080	0.173	-3.320	0.146	-0.018	0.809
	Response Order	5.122	0.000	-1.080	0.597	3.240	0.113	2.000	0.306	-0.140	0.021
	Odd/even	12.191	0.000	0.920	0.540	0.600	0.687	-0.800	0.618	0.12	0.179
	Opinion Float	2.444	0.000	0.920	0.540	0.600	0.687	-0.800	0.618	-0.033	0.714
	Allow/forbid	-42.200	0.000	-6.200	0.004	2.250	0.332	0.350	0.877	-0.628	0.000
Llama2 7b-chat	Acquiescence	1.136	0.647	-7.807	0.000	-12.034	0.000	-5.546	0.000	-0.099	0.189
	Response Order	-9.801	0.000	7.173	0.000	12.679	0.000	1.594	0.253	-0.315	0.000
	Odd/even	20.079	0.000	8.460	0.000	15.810	0.000	9.175	0.000	-0.315	0.000
	Opinion Float	-1.254	0.283	8.460	0.000	15.810	0.000	9.175	0.000	-0.086	0.339
	Allow/forbid	-7.050	0.367	-18.700	0.000	-24.600	0.000	-16.200	0.002	-0.161	0.321
Llama2 13b-chat	Acquiescence	1.909	0.434	-9.239	0.000	-11.534	0.000	-5.284	0.000	-0.095	0.209
	Response Order	-9.292	0.000	7.653	0.000	10.753	0.000	0.472	0.719	-0.324	0.000
	Odd/even	21.254	0.000	10.159	0.000	14.460	0.000	9.492	0.000	-0.163	0.069
	Opinion Float	-0.191	0.870	10.159	0.000	14.460	0.000	9.492	0.000	-0.106	0.238
	Allow/forbid	-7.300	0.333	-15.950	0.000	-23.450	0.000	-16.200	0.000	-0.131	0.422
Llama2 70b-chat	Acquiescence	11.114	0.000	2.320	0.523	-5.280	0.312	4.040	0.166	0.452	0.000
	Response Order	-0.495	0.745	0.200	0.904	15.040	0.002	1.200	0.459	0.465	0.000
	Odd/even	26.476	0.000	3.280	0.210	-2.040	0.656	-7.240	0.018	-0.231	0.009
	Opinion Float	1.556	0.039	3.280	0.210	-2.040	0.656	-7.240	0.018	0.440	0.000
	Allow/forbid	4.000	0.546	-4.750	0.258	-16.000	0.021	-0.950	0.811	0.280	0.080
Solar	Acquiescence	18.511	0.000	0.120	0.970	2.560	0.596	0.600	0.833	0.187	0.013
	Response Order	9.683	0.000	2.280	0.336	8.680	0.012	4.360	0.017	0.248	0.000
	Odd/even	17.508	0.000	0.480	0.815	2.960	0.223	-1.000	0.661	-0.385	0.000
	Opinion Float	1.921	0.017	0.480	0.815	-2.960	0.223	-1.000	0.661	0.291	0.001
	Allow/forbid	6.800	0.207	-2.950	0.343	-8.500	0.131	-8.050	0.001	0.145	0.373
GPT 3.5 Turbo	Acquiescence	5.523	0.040	-11.720	0.008	-28.680	0.000	-19.120	0.000	0.334	0.000
	Response Order	-2.709	0.147	4.960	0.121	15.960	0.002	8.000	0.011	0.198	0.001
	Odd/even	25.048	0.000	-5.480	0.082	-14.800	0.001	-5.800	0.062	-0.273	0.002
	Opinion Float	-11.905	0.000	-5.480	0.082	-14.800	0.001	-5.800	0.062	0.467	0.000
	Allow/forbid	25.300	0.000	-12.000	0.008	-23.200	0.001	-6.950	0.058	0.206	0.202
GPT 3.5 Turbo Instruct	Acquiescence	6.455	0.024	2.600	0.445	-11.800	0.008	-2.800	0.326	0.334	0.000
	Response Order	-11.114	0.000	3.880	0.169	11.920	0.001	3.800	0.147	0.275	0.000
	Odd/even	2.032	0.390	1.560	0.433	-7.120	0.061	-0.840	0.711	-0.073	0.416
	Opinion Float	0.143	0.891	1.560	0.433	-7.120	0.061	-0.840	0.711	0.360	0.000
	Allow/forbid	8.550	0.111	-4.500	0.216	-10.050	0.139	4.100	0.261	0.437	0.005

Table 4:  $\bar{\Delta}_b$  for each bias type and associated p-value from t-test as well as  $\bar{\Delta}_p$  for the three perturbations and associated p-value from t-test. We also report the Pearson r statistic between model uncertainty and the magnitude of  $\Delta_b$ .