Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

James Bisbee* Joshua D. Clinton† Cassy Dorff‡

Brenton Kenkel§ Jennifer M. Larson¶

August 9, 2023

Abstract

Large Language Models (LLMs) offer new research possibilities for social scientists, but their potential as "synthetic data" is still largely unknown. In this paper, we investigate how accurately the popular closed-source LLM ChatGPT can recover public opinion, prompting the LLM to adopt different "personas" and then provide feeling thermometer scores for 11 sociopolitical groups. The average scores generated by ChatGPT correspond closely to the averages in our baseline survey, the 2016–2020 American National Election Study. Nevertheless, sampling by ChatGPT is not reliable for statistical inference: there is less variation in responses than in the real surveys, and regression coefficients often differ significantly from equivalent estimates obtained using ANES data. We also document how the distribution of synthetic responses varies with minor changes in prompt wording, and we show how the same model yields significantly different results over a three-month period. Altogether, our findings raise serious concerns about the quality, reliability, and reproducibility of synthetic survey data generated by LLMs.

Word Count: 9,884 (excluding title page and text on figures)
Keywords: ChatGPT, synthetic data, public opinion, research ethics

*Assistant Professor of Political Science, Vanderbilt University james.h.bisbee@vanderbilt.edu. Corresponding author.
†Abby and Jon Wikelried Professor, Vanderbilt University josh.clinton@vanderbilt.edu
‡Assistant Professor of Political Science, Vanderbilt University cassy.dorff@vanderbilt.edu
§Associate Professor of Political Science, Vanderbilt University brenton.kenkel@vanderbilt.edu
¶Associate Professor of Political Science, Vanderbilt University jennifer.larson@vanderbilt.edu

1 Introduction

Public opinion polling is seemingly in a crisis (Shapiro 2019). Costs are increasing, response rates are declining (Keeter et al. 2017), and there are growing concerns about inaccuracy (Kennedy et al. 2018; Clinton et al. 2021a, 2022). These patterns have raised questions about biases due to coverage and non-response (e.g., Cavari and Freedman 2022, though see Mellon and Prosser 2021) and have even led some to question the viability of polling itself (Meyer et al. 2015; Keeter 2018). Assessments of public opinion based only on the most accessible and numerous groups risk missing important voices in an ever more diverse polity (Brehm 1993; Berinsky 2004). At the same time, polls are necessary tools to assess and address growing concerns about polarization and democratic backsliding (Graham 2023; Waldner and Lust 2018). When done well, public opinion polls allow scholars, policymakers, and journalists to assess the opinions of the whole public—not just those who are the most active, willing, and able to express their opinions through more costly means like direct appeals, protests, and donations.

Given the rising expense and difficulty of interviewing respondents, researchers increasingly turn to other methods of characterizing public opinion and sentiment—especially for groups that are smaller or harder to reach (Wang et al. 2015; Ghitza and Gelman 2020; van Klingeren et al. 2021). Some use non-survey data, most often from social media platforms like Twitter, to characterize public opinion (e.g., Beauchamp 2017; Tucker 2017; though see Bail 2022). Others use sophisticated weighting methods like multilevel regression poststratification to leverage observed relationships among respondents who are surveyed to characterize the opinions of those who are not (Gelman 1997; Lax and Phillips 2009; Ghitza and Gelman 2013; Caughey and Warshaw 2015; Bisbee 2019; Goplerud 2023).

Large Language Models (hereafter LLMs) that synthesize vast corpora of human-generated text—including political news, opinion, and debate— might look like the new frontier in characterizing public opinion without the expense of traditional polling. Nearly every day, a new working paper asserts the benefits and possibilities of this technology. Social scientists have already used LLMs to label political data (Törnberg 2023; Gilardi et al. 2023), estimate politicians' ideology (Wu et al. 2023), and generate synthetic samples for pilot testing (Argyle et al. 2023; Horton 2023). At least one startup in private industry suggests that "synthetic users" can supplement or replace

human respondents in development and marketing.[1] If LLMs have the potential to reorient nearly every facet of economic and social life (Bommasani et al. 2022), then perhaps they could revolutionize public opinion research too. Can a pretrained LLM produce synthetic opinions for respondent personas that accurately mirror what similar human respondents would say on a real survey?[2] Instead of spending $14 million to interview humans in the 2024 American National Election Study, for example, might researchers instead spend $90 (the cost of the initial study we conducted) to obtain an accurate portrayal of American public opinion using synthetic samples from an LLM?

To evaluate these questions, we prompt ChatGPT 3.5 Turbo[3] (OpenAI 2021) to first adopt various personas defined by demographic and political characteristics [4] and then answer a battery of questions about feelings towards social and political groups. We refer to these responses as "synthetic data", which is occasionally referred to in related research as "silicon samples" (Argyle et al. 2023).[5] To facilitate comparison to widely used public opinion data, the characteristics defining each persona in our synthetic data are taken from real respondents in the 2016 and 2020 American National Election Study, and our survey questions closely mirror the ANES's feeling thermometer questions.

Our primary analysis compares the distribution of responses from synthetic ChatGPT personas to matching corresponding respondents in the ANES. We focus on three metrics of interest to social scientists: (1) how well ChatGPT recovers the overall mean and variance of feelings towards various groups, (2) how closely the (conditional) correlations between persona characteristics and

[1] https://www.syntheticusers.com.

[2] We rely on the corpus contained in a pretrained LLM rather than fine-tuning (e.g., as in Argyle et al. 2023) because this workflow is more accessible and more likely to be used by journalists, politicians, and the modal academic (Cowen 2022).

[3] We focus on ChatGPT 3.5 Turbo because its release was what prompted the flurry of researcher and public interest that initially motivated this project. More practically, it was the largest and most popular pretrained LLM with public API access at the time we began our research. We replicate some of our results using the even larger ChatGPT 4.0 and the open-source model Falcon-40B-Instruct; see SI Sections 11 and 14 for details.

[4] By prompting the LLM to adopt particular characteristics, including party identification and political ideology, we differ from earlier research identifying political bias in the "default" persona (Rozado 2023; Santurkar et al. 2023). In an auxiliary analysis, we find similar results on ChatGPT's default bias; see SI Section 7.

[5] Our approach differs in the details from Argyle et al. (2023) who fine tune a generative pre-trained transformer (GPT) using real survey data to generate their "silicon" samples. In contrast, we simply ask ChatGPT 3.5 to adopt a given persona to collect our synthetic data.

survey responses mirror the inferences we would draw from the ANES, and (3) the sensitivity of our comparisons to changes in the prompt, the LLM, and the timing of data collection. To briefly summarize, ChatGPT produces feeling thermometer scores whose averages closely track baseline ANES values, but it performs poorly by our other metrics—including the ones that represent standard inferential tasks in political science research.

At the coarsest level of analysis, synthetic ChatGPT opinions look remarkably similar to human ANES respondents. Across the 11 groups for which we generate feeling thermometer scores, the average response in our synthetic data is always within a standard deviation of the ANES average (range of 5.2 to 28.9 standard deviations for the synthetic data, and 19.5 to 31.9 for the ANES data). However, even when we compare overall average responses, we find problems with how well ChatGPT recovers the distribution of public opinion. The standard deviation of each feeling thermometer in our synthetic data is on average about half (0.62) as large as in the ANES data and in 4 out of 11 cases, the standard deviation is less than half the ANES value. Even if the averages from the synthetic sample provide a roughly accurate high-level look at public feelings towards different groups, the underestimation of variance sharply limits the LLM's usefulness for statistical inference, including for the purpose of pre-testing questions or piloting studies.

The synthetic sample fares worse when we examine higher-order relationships. First, while our overall averages closely correspond to the ANES values, the same is not true when we investigate substantively relevant subgroups. For example, compared to the subgroup averages in the ANES data, the synthetic estimates overestimate Democrats' evaluation of liberals and the Democratic Party, while underestimating their feelings toward conservatives. The problem of insufficient variation is even more severe at the subgroup level; the standard deviation of opinions for a given persona profile is on average just 31% that of matched human opinion from the ANES data. Second, when we regress feeling thermometer scores on respondents' demographic attributes—the type of analysis common in public opinion research—the synthetic sample would frequently lead us to draw different inferences than if we relied on human respondents. 48% of coefficients estimated from the ChatGPT responses are statistically significantly different from their ANES-derived counterpart; among these cases, the sign of the effect flips 32% of the time. Simply put, a political scientist who runs a public opinion regression on synthetic responses from a pretrained LLM cannot have confidence that their effect estimates will be qualitatively similar to what they would obtain from

a traditional survey.

If these statistical mismatches were not enough of a problem on their own, we document auxiliary issues leading us to question the reliability and reproducibility of sampling via ChatGPT. The most concerning of these is that the distribution of responses to the same prompt changed between our initial run in April 2023 and a rerun in July 2023, even though we were ostensibly sampling from the same ChatGPT 3.5 Turbo model. This is a key illustration of how closed-source generative models pose a threat to the reproducibility norms of contemporary social science (Spirling 2023). Besides these reproducibility concerns, we also find that the distribution of results is sensitive to small differences in the prompt structure: For example, the means of the synthetic thermometers come closer to their ANES counterparts when we use a first-person rather than a third-person prompt, even though the same demographic information is presented to the LLM in both cases.

Our findings raise serious questions about the use and performance of LLMs for the characterization of human opinion and the creation of synthetic data, connecting with a growing body of research that asks similar questions (Bender et al. 2021; Spirling 2023) and documents similar issues (Motoki et al. 2023; Rozado 2023; Abdulhai et al. 2023; Cao et al. 2023). When we prompt LLMs to adopt personas matching actual ANES respondents, we end up with data that largely fails to replicate our best estimates of the correlates of human opinion.[6] While it is, of course, possible that continuing technological innovations will someday produce more stable, knowable, and accurate synthetic responses than the ones we document, our work challenges the utility of using LLMs to create synthetic data for social science research.

2 Research Design and Data

An LLM is a prediction algorithm optimized to predict the next token in a sequence of text data. When prompted to take on a persona with a set of attributes and answer a question from that perspective, contemporary LLMs can provide remarkably coherent responses. The largest models have demonstrated remarkable emergent abilities well outside the core text-sequencing tasks at

[6] It is possible that the ANES is worse at recovering the true population parameters of interest than the LLM, perhaps due to the issues with current public opinion polling discussed above. However, without a census of public opinion against which to compare, it seems irresponsible to jump to this interpretation of the discrepancies we document below between the synthetic and human samples.

the core of their training (Wei et al. 2022). The uncanny coherence of LLM responses to a wide variety of prompts have generated considerable excitement about the possibility of using these models to generate responses that are representative of public opinion. For example, Argyle et al. (2023) conclude that "by conditioning the model on thousands of socioeconomic backstories from real human[s]," LLMs are able to generate synthetic opinions that "reflect[] the complex interplay between ideas, attitudes, and sociocultural context that characterize human attitudes."[7]

But how far can we push this conclusion? If we ask an LLM to act as though it were a 30-year old White male Republican with a high-school degree and ask about its feelings toward Democrats, can we expect its responses to mirror the distribution of opinions in that group?[8] The vast corpus used to train these models (see, e.g., Washington Post 2023) contains reams of political writing that could, perhaps, be used to construct nuanced depictions of opinions across groups. But when that vast corpus comes from the Internet, the content may be unrepresentative of the public at large, or at least some groups within it (Bail 2022).[9] Scholars have previously probed the default persona of ChatGPT by prompting it to answer a battery of survey questions and showing that the resulting responses have ideological, dispositional, and psychological biases (Motoki et al. 2023; Rozado 2023; Bail 2022; Abdulhai et al. 2023), but it is unknown how well pretrained LLMs can reproduce human opinion when prompted to adopt a particular persona. The precise recipe that generates responses from ChatGPT is a trade secret,[10] so our approach is to generate a data set of synthetic opinions from ChatGPT that we can compare to opinions from a survey of human respondents.

To examine a pretrained LLM's ability to simulate the opinions of respondents defined by a particular set of features, we use a survey instrument known as a "feeling thermometer." Respon-

[7] Argyle et al. (2023) use GPT-3, a predecessor to the GPT-3.5 LLM underlying the ChatGPT program we query in our main analysis.

[8] Our main analysis focuses on whether we *can* do this. A more fundamental question that we return to in the conclusion is whether we *should*.

[9] Volumes of research have shown that humans online are less thoughtful and more hurtful (Lapidot-Lefler and Barak 2012; Rowe 2015); less able to reason and quicker to rely on stereotypes and heuristics (Halpern and Gibbs 2013); more willing to villainize those who disagree with them and less willing to engage in empathy (Rowe 2015; Rossini 2022); and more inclined to express themselves to signal group attachments (Bail 2022).

[10] For this reason, some scholars have argued for rejecting proprietary, closed-source LLMs (Spirling 2023)—an argument that prompted our investigation of the open-source Falcon-40B-Instruct model (see SI 14).

dents are instructed to consider some group and to indicate the degree to which they experience warm (positive, affectionate, etc.) or cool (negative, disdainful, etc.) feelings toward members of that group. This instrument has been employed by major polling outfits since 1964, and it is widely used by scholars to characterize the extent of societal division (Druckman and Levendusky 2019; Iyengar and Westwood 2019). After instructing the LLM to adopt a persona defined by a set of respondent characteristics, we then ask it to answer a battery of feeling thermometer questions. We choose respondent personas and feeling thermometers to match questions that human respondents were asked in a real public opinion survey, allowing us to gauge how accurately and precisely the synthetic LLM responses capture actual human opinion.[11]

We conduct our primary investigation using perhaps the most publicized and popular software built on a pretrained LLM, ChatGPT 3.5 Turbo (OpenAI 2021). To characterize how robust our results are to differences across LLMs (as well as changes in the "same" LLM over time), we replicate portions of our analysis with ChatGPT 4.0; with ChatGPT 3.5 Turbo both before and after its update on June 25, 2023; and with the open-source model Falcon-40B-Instruct.[12]

We iterated on several different versions of a prompt to gather our synthetic data, the full description of which can be found in the SI Section 1. Our main results are based on the most detailed version of a prompt in which we instructed ChatGPT to adopt a persona defined as follows:

> "It is `[YEAR]`. You are a `[AGE]` year-old, `[MARST]`, `[RACETH]` `[GENDER]` with `[EDUCATION]` making `[INCOME]` per year, living in the United States. You are `[IDEO]`, `[REGIS]` `[PID]` who `[INTEREST]` pays attention to what's going on in government and politics."[13]

In each query to ChatGPT, the characteristics in brackets were substituted with values corresponding to a real respondent in the 2016 or 2020 wave of the ANES. These values include:

[11] We also test a battery of other questions that rely on conventional Likert-style survey instruments, pertaining to the role of government in the economy, and to restrictions placed on protesters and revolutionaries. These results can be found in SI Section 12.

[12] Highlighting issues with replicability, our original analyses were conducted on the pre-June 25 version of ChatGPT 3.5 Turbo, but that version is only accessible until September 2023.

[13] We chose these covariates based on their utilization for weighting and poststratification in the survey methods literature (Lax and Phillips 2009; Ghitza and Gelman 2013), as well as adhering to the norms of weight calculation in the ANES itself (DeBell and Krosnick 2009). In Section 4.1, we demonstrate that the synthetic data is sensitive to which attributes are included in the prompt, finding that the majority of the performance hinges on the political covariates.

- [YEAR]: 2016 or 2020
- [AGE]: age in years of ANES respondent
- [RACETH]: non-Hispanic white, non-Hispanic black, or Hispanic
- [GENDER]: male or female
- [MARST]: divorced, married, separated, single, or widowed
- [EDUCATION]: a high school diploma, some college but no degree, a bachelor's degree or more
- [INCOME]: $30k, $50k, $80k, $100k, $150k or more
- [IDEO]: an extremely liberal, a liberal, a slightly liberal, a moderate, a slightly conservative, a conservative, an extremely conservative
- [REGIS]: registered, unregistered
- [PID]: Democrat, Independent, Republican
- [INTEREST]: never, sometimes, frequently, regularly, always

For our primary analyses, the LLM was then prompted as follows:

> "Provide responses from this person's perspective. Use only knowledge about politics that they would have. Format the output as a tsv table with the following format:
>
> group,thermometer,explanation,confidence
>
> The following questions ask about individuals' feelings toward different groups. Responses should be given on a scale from 0 (meaning cold feelings) to 100 (meaning warm feelings). Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group. You would rate the group at the 50 degree mark if you don't feel particularly warm or cold toward the group.
>
> How do you feel toward the following groups?
> The Democratic Party?
> The Republican Party?
> Democrats?
> Republicans?
> Black Americans?
> White Americans?
> Hispanic Americans?
> Asian Americans?
> Muslims?
> Christians?
> Immigrants?
> Gays and Lesbians?
> Jews?

Liberals?
Conservatives?
Women?"

We prompt 30 synthetic respondents for each of the 7,530 human respondents in the ANES survey, yielding a final dataset of 3,614,400 responses.[14] For each response, we record both the numeric feeling thermometer score, the explanation provided by the LLM for why they chose the score, and a measure of the model's reported confidence in the response.[15] Even though we used the cheaper ChatGPT 3.5-turbo for creating this synthetic dataset, the overall cost was roughly $600 US dollars. In our main analyses below, we rely on the average response from the 30 synthetic respondents drawn for each human.

3 Results

We evaluate the synthetic data's quality in three ways. We begin by simply comparing means and variances to demonstrate that although ChatGPT appears to perform reasonably well at recovering overall means, closer inspection reveals that it is often biased and overly confident in its approximations of real human survey responses. We then test whether ChatGPT can recover marginal associations between covariates that might be of interest to researchers and find that regression results estimated using human and synthetic samples often differ, sometimes substantially in systematic ways. Finally, we document considerable sensitivity of synthetic responses to both the timing of when a prompt was used to generate data, and also to the persona each prompt represents.

[14] Data collection issues prevented us from getting exactly 30 synthetic respondents for each person (e.g., when ChatGPT would not format the data correctly). The vast majority of human respondents were paired with exactly 30 synthetic respondents, and each human respondent was paired with at least 10. We find that these errors were significantly more likely for political groups (the political parties, liberals and conservatives, and gays & lesbians); and for personas that are less educated, unregistered, white, Democrat, male, and less wealthy. However, none of these differences amount to more than one or two missing synthetic samples out of 30. Importantly, we found little evidence of these errors being the product of the AI refusing to provide an answer, but were instead due to incorrectly formatted tsv results.

[15] We use the explanation and the confidence for validation tests of our results presented in our Supporting Information Section 5, testing whether the explanations for a given temperature cohere with the score chosen, and whether the AI reports lower confidence for certain target groups or certain personas than others.

We underscore that we selected our empirical setting as a theoretically easy test for ChatGPT.[16] Existing work has documented a western, especially American, bias in the LLM (Cao et al. 2023); we benchmark using one of the most well-known American political opinion surveys; and we specify the precise year that the human respondent participated in the survey, a year prior to the end of ChatGPT's training period. Even so, we test the sensitivity of our conclusions to other questions and datasets in the Supporting Information, finding significantly worse performance than we document in this paper, and supporting our intuition that what follows is something of a best-case scenario for synthetic data. Yet even in this best-case scenario, the best we can say is that the overall average synthetic responses are close to the population averages. For the kinds of associational questions that social scientists care about, our findings throw cold water on the idea that researchers can substitute synthetic survey data for human respondents.

3.1 Accuracy of Average and Standard Deviations

Figure 1 begins by plotting the sample means and standard deviations for the various feeling thermometer scores estimated using either the human respondents to the ANES or the average of the synthetic respondents drawn from ChatGPT.[17] Because the synthetic data were collected to exactly match ANES respondents on the selected covariates, the distributions ideally would be identical.

While the average of ChatGPT responses does not exactly match the average survey response in the ANES, every synthetic mean falls within one standard deviation of the ANES average. In addition, the rank ordering of feeling thermometers is largely intact across both samples. That said, the distribution of synthetic responses for some questions exhibit far less variation than human responses, especially for questions asking about feelings toward racial and religious groups. The variation in the ChatGPT responses also reflects statistical uncertainty due to sampling from the underlying language model,[18] so it is striking that these responses are still more tightly distributed

[16] Given that ChatGPT has far more parameters than competing models, and LLM performance is known to increase with model size to date (Wei et al. 2022), ChatGPT itself is an easy test for the broader question of how well pretrained LLMs can generate synthetic data.

[17] Five of our 16 target groups were not asked in the ANES waves: Hispanic Americans, Democrats, Republicans, women, and immigrants, although were included in the alternative dataset that we examine in SI Section 6.2.

[18] We use a temperature parameter of 1 in our main analysis, and demonstrate the strong positive association

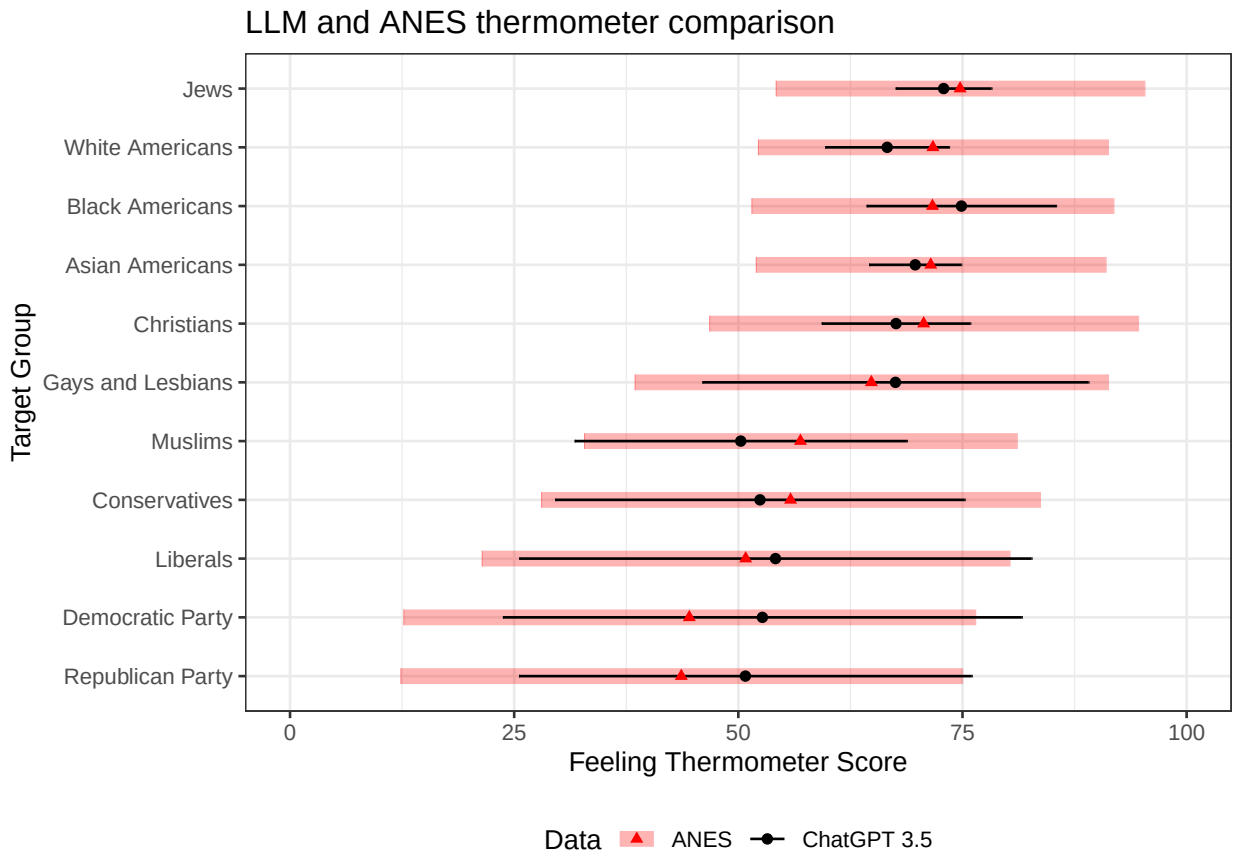

Figure 1: Average feeling thermometer results (x-axis) for different target groups (y-axis) by prompt type / timing (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical.

than in the ANES data.

Even though the synthetic data produced by ChatGPT broadly performs well in terms of summarizing overall human opinion, issues emerge when we look at subgroups. To demonstrate, we examine affective polarization and partisan sectarianism, calculating how average opinions toward liberals, conservatives, and the major parties vary across groups of respondents defined by race and partisanship. Figure 2 presents the results, highlighting the relative extremity of ChatGPT responses, especially among Democrats, that was masked when averaging over partisanship in Figure 1. These differences are substantively meaningful, amounting to 0.5–1 standard deviations of the ANES distribution of attitudes, and 10–20 points on the 100-point thermometer scale. In

between this hyperparameter and the empirical variance of the synthetic data in SI Section 2.

particular, these results suggest that Democrats like liberals more, and conservatives less, than their human counterparts, exaggerating the out-group antipathy along ideological lines. Similar extremism is found among Republicans, especially among non-Hispanic Black Republicans. While differences in the synthetic and actual responses of independents sometimes occur, there is less systematic evidence of exaggerated polarization in terms of partisanship. In general, the patterns reported in Figure 2 highlight that synthetic responses would suggest that society is more politically hostile than it actually is.

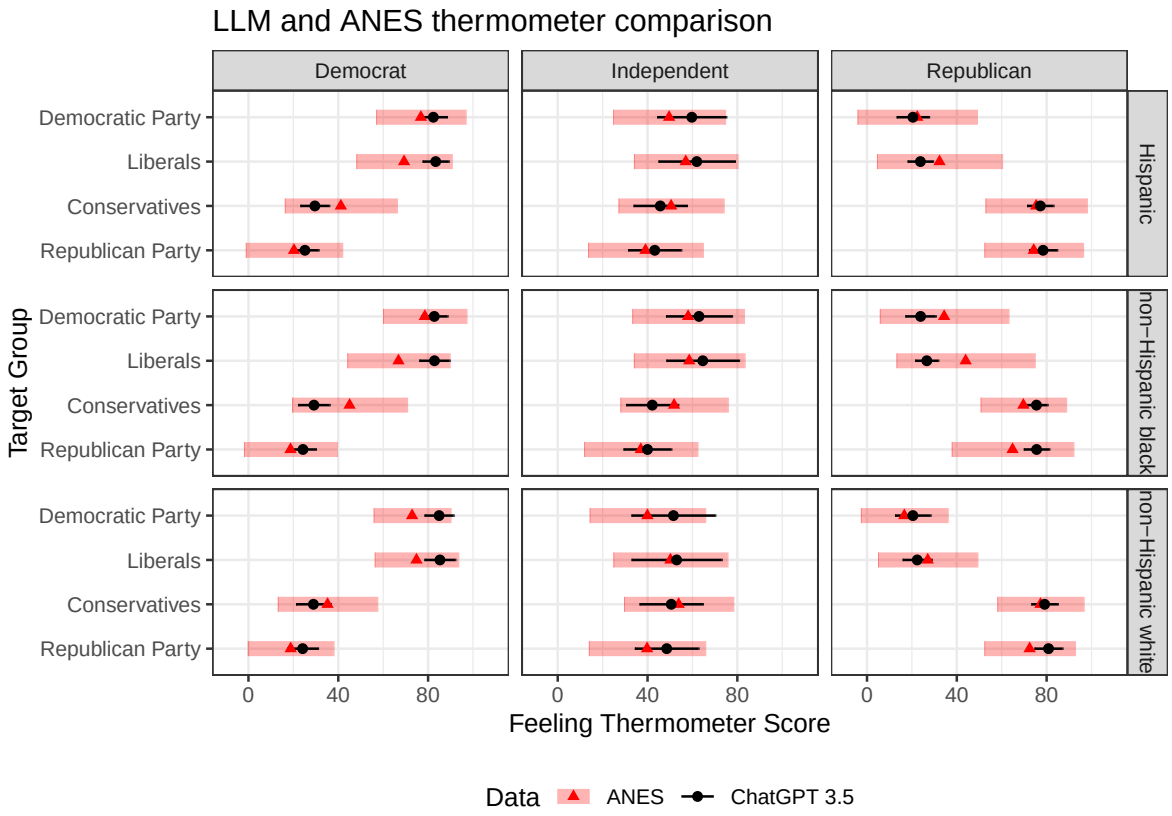

Figure 2: Average feeling thermometer results (x-axis) for different target groups (y-axes) by party ID of respondent (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated by black circles and thin black bars. Sample sizes for each group-wise comparison are identical.

Figure 2 also reveals far smaller standard deviations in the synthetic estimates than found in the ANES—ChatGPT does not just exaggerate partisan differences, it is also more confident in this conclusion. Beyond the substantive concerns with this over-confidence, the undersized variance of synthetic responses poses serious inferential problems for attempts to use such data for pre-analysis

study design as some scholars have suggested (Argyle et al. 2023). Consider, for example, using synthetic opinions from ChatGPT to conduct a power analysis for a human-respondent test of whether partisan affective polarization has increased since 2012, when the average gap between in-party and out-party assessments among partisans in the ANES was 47.4. Table 1 reports the results. Using the estimates of the magnitude and variation in the ChatGPT-generated measures of affective polarization, we calculate the sample size required to detect a difference from the 2012 level at various levels of power. As a baseline, we perform the same power calculation using the magnitude and variation in feeling thermometer scores from our ANES comparison set. Even for 99% power, the ChatGPT estimates imply that just 33 partisan respondents would be necessary to detect a difference from affective polarization in 2012—an underestimation that is almost an order of magnitude less than what we calculate from the ANES benchmark.[19]

	Sample Size Needed	
Power	ANES est.	ChatGPT est.
80%	129	15
85%	147	17
90%	172	20
95%	212	24
99%	299	33

Table 1: Calculations of the sample size necessary for a specified power to reject the null hypothesis of no difference in affective polarization among partisans from the average level in the 2012 ANES, assuming a 95% significance level. The second column records the calculation if we assume an effect size and variance equal to the 2016–2020 pooled ANES values (size 7.8, sd 31.4); the third column is the same calculation with our ChatGPT estimates (size 12.5, sd 16.1).

[19] Here the problem arises from the mismatch of using synthetic responses from a LLM to to guide the design of a study of human respondents. Of course it is possible that the LLM is providing a good estimate of *something else* with appropriate variation, but we leave that deeper point for future research. Our more limited point is that the over-precision of the LLM makes it a possibly misleading tool for guiding the design of human-respondent studies. We include a detailed analysis of variability of the synthetic and human samples in SI Section 9.

3.2 Accuracy of Estimated Regression Coefficients

Although synthetic data sometimes looks reasonably accurate—albeit too precise—in the aggregate, more concerning problems emerge when looking at conditional relationships. To test whether the correlational structure of the synthetic data corresponds to the ANES benchmark, we examine differences in regression results obtained using true and synthetic feeling thermometers as the dependent variable. This is a critical test for political science, where scholars are typically interested in the correlates of public opinion rather than in broad average values.

We estimate linear regression models of the following form:

$$\text{FT}_{i,d} = \alpha + \gamma \mathbb{I}_d + \boldsymbol{\beta} \cdot \mathbf{x}_i + \boldsymbol{\lambda} \cdot (\mathbb{I}_d \mathbf{x}_i) + \epsilon_{i,d}, \tag{1}$$

where i indexes respondents, d indexes data source (ANES or ChatGPT), $\mathbb{I}_d$ is an indicator for the data source being ChatGPT, and $\mathbf{x}_i$ is a vector of the respondent characteristic used in our persona prompt (age, gender, race, education, income, marital status, ideology, party ID, voter registration status, and interest in news and politics). We are most interested in the vector of $\boldsymbol{\lambda}$ coefficients measuring how the partial correlation between each covariate in $\mathbf{x}_i$ and the feeling thermometer score differs between the synthetic responses and the ANES benchmark.

We run the specification in Equation 1 for each combination of feeling thermometer (the 11 target groups in Figure 1) and survey year (2016 or 2020), for 22 total regressions. Figure 3 plots the resulting coefficient estimates estimated using the ANES (x-axis) versus the synthetic data (y-axis), broken out by the predictor. Points are shaded based on whether the coefficients estimated on the synthetic data are significantly different from those estimated on the ANES. Points closer to the 45 degree line indicate a better correspondence between the conclusions an applied researcher would draw using either human or synthetic datasets. Points in the off-diagonal quadrants (upper left and bottom right) are coefficients whose sign differs depending on which data source we employ.

The plot highlights the degree to which the synthetic data is influenced by political covariates. The ideology measure performs best, with points lying close to the 45 degree line, many of which do not differ significantly between the datasets. Partisanship also exhibits a strong positive association, although the evidence of an S-shape suggests that the associations between partisanship and the battery of feeling thermometers are stronger in the synthetic data than in the human data.

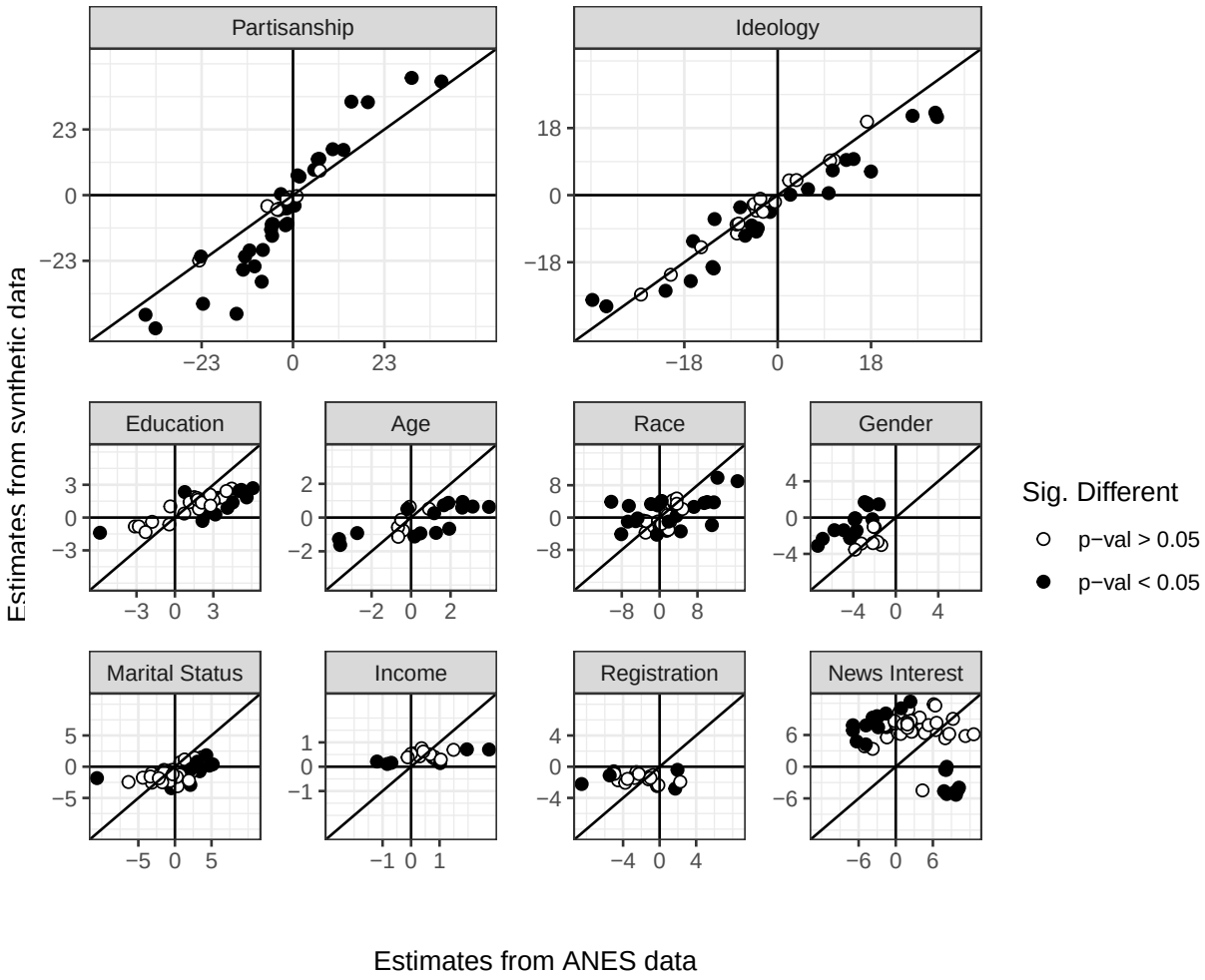

Figure 3: Each point describes the coefficient estimate capturing the partial correlation between a covariate and a feeling thermometer score toward one of the target groups, estimated in either 2016 or 2020. The x-axis position is the coefficient estimated in the ANES data, and the y-axis position is the same coefficient estimated in the synthetic data. Solid points indicate coefficients who are significantly different when estimated in either the ANES or synthetic data, while hollow points are coefficients that are not significantly different. Points in the northeast and southwest quadrants generate the same substantive interpretations, while those in the northwest and southeast quadrants produce differing interpretations. A synthetic dataset that is able to perfectly recover relationships estimated in the ANES data would have all points falling along the 45 degree line.

Conversely, we document far worse performance along other covariates, in several instances leading to substantively different conclusions—sometimes with opposite signs—than what we would learn from the actual ANES.

One potential explanation for these results is that the synthetic data is overdetermined by the

political covariates. We demonstrate this point with an analysis of age, comparing and contrasting the results of a bivariate and multiple regression specification run on the ANES and synthetic datasets. The bivariate specification predicts feeling thermometer scores toward the two major political parties as a function of age:

$$\text{FT}_{i,d} = \alpha + \gamma \mathbb{I}_d + \beta \text{age}_i + \lambda(\mathbb{I}_d \times \text{age}_i) + \epsilon_{i,d}. \tag{2}$$

We compare the estimated marginal effect of age using the ANES data (β) and using the ChatGPT responses ($\beta + \lambda$) from this bivariate specification to the corresponding values from the controlled specification estimated in Equation 1. As above, we run separate regressions for the Democratic and Republican Party feeling thermometers in both 2016 and 2020. We plot the marginal effects in Figure 4, revealing that not only does the synthetic data produce significantly different relationships than the ANES data in the bivariate specification, it also flips the sign for attitudes towards the Democratic Party in the specification that includes controls.

The comparison of the bivariate and multivariate specifications is illuminating. While sampling from ChatGPT can recover basic relationships between age and feelings towards the major American parties, it does not accurately encode the conditional associations that result in a positive relationship between FT scores and age, all else equal. This shortcoming can be articulated in the language of omitted variable bias. Age is correlated with several other characteristics that predict GOP support in the United States, such as income (Angel and Settersten Jr 2013), marital status (Angel and Settersten Jr 2013; Glenn 1974), and ideological conservativism (Gerber et al. 2010; Cornelis et al. 2009). These omitted variables generate the spurious conclusion that age "causes" warmer feelings towards the Republican Party and cooler feelings toward the Democratic Party.[20] In survey data generated by real humans, conditioning on these other characteristics holds them constant, revealing that—all else equal—older respondents are warmer toward *both parties* compared to younger Americans. Yet the synthetic data is unable to achieve this degree of nuance, instead reiterating the spurious conclusion of the bivariate regression: the older you are, the less

[20]The association between age and conservativism is so wide-spread that it is accompanied by the aphorism that a conservative 20 year-old doesn't have a heart, while a liberal thirty year-old doesn't have a brain, (Peterson et al. 2020).

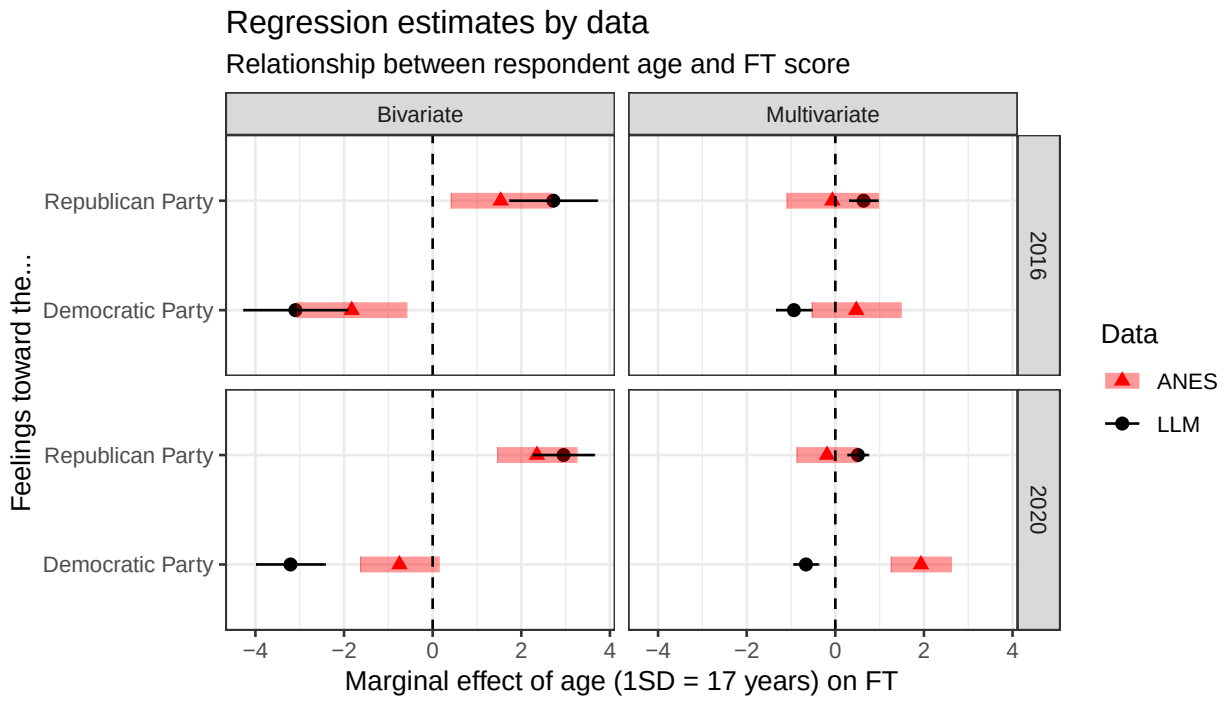

Figure 4: Coefficients (points) and 95% confidence intervals (bars) describing the relationship between age and feelings toward the major American political parties, broken out by year (rows) and specification (columns). Results estimated in the ANES data are indicated with red triangles and large transparent red bars, while the same results estimated in the synthetic data are indicate with black circles and narrow black bars. The bivariate specification (left column) predicts FT scores only as a function of age, concluding that older respondents are more positively oriented toward the Republican party and more negatively oriented toward the Democratic party, a conclusion that is recovered (albeit exaggerated) in the synthetic data. The multivariate specification (right column) estimates the same relationship, controlling for all other covariates used to describe the persona to the LLM. While the ANES data indicates no relationship between age and attitudes in 2016 and, if anything, a positive association for the Democratic party in 2020, the synthetic data continues to produce the same association documented in the bivariate specification.

warm you feel toward the Democratic Party.

4 Sensitivity of Synthetic Responses

Even in the best-case scenario, synthetic data from ChatGPT has too little variation across respondents and fails to recover substantively important conditional relationships. In the course of our research establishing these problems, we uncovered additional concerns about the reliability and replicability of synthetic sampling via ChatGPT. Specifically, we find that the distribution of responses is highly sensitive to differences in the prompt used to generate data, which version of

ChatGPT is used, and even changes over time in the "same" model.

4.1 Effect of Different Prompts

How sensitive are our findings to different prompt specifications? To investigate, we re-collected the synthetic data with two modifications to the description of the persona. The first prompt only described the basic demographic profile of the synthetic respondent, including their age, gender, race, marital status, education, and income. The second prompt only included a description of the synthetic respondent's political characteristics, including their ideology, partisanship, registration status, and interest in news and politics. The full prompt (the basis of the preceding results) included all attributes in a single description. The raw code for these prompts is given below, where the placeholder text is indicated with a capitalized characteristic in square brackets, which would be replaced with a description for the relevant ANES respondent as described above:

- Demographics only: "It is `[YEAR]`. You are a `[AGE]` year-old `[MARST]` `[RACETH]` `[GENDER]` with `[EDUCATION]` making `[INCOME]` per year, living in the United States."
- Politics only: "It is `[YEAR]`. You are `[IDEO]`, `[REGIS]` `[PID]` who `[INTEREST]` pays attention to what's going on in government and politics, living in the United States."
- Combined: "It is `[YEAR]`. You are a `[AGE]` year-old, `[MARST]`, `[RACETH]` `[GENDER]` with `[EDUCATION]` making `[INCOME]` per year, living in the United States. You are `[IDEO]`, `[REGIS]` `[PID]` who `[INTEREST]` pays attention to what's going on in government and politics."

We investigate the sensitivity by predicting the mean absolute error (the absolute difference between the ANES feeling thermometer and the LLM's estimate) as a function of the prompt interacted with the target group and the party ID, controlling for all other covariates. As illustrated in Figure 5, the average absolute error is basically identical between the full prompt and politics-only prompt. However, failing to include information on the respondent's politics dramatically inflates the error for certain groups, notably those groups which are more politically salient (the parties, ideological groups, gays and lesbians, and Muslims). Political descriptions do not dramatically

change the LLM's errors when it comes to predicting the feeling thermometers toward racial or religious groups, except for Muslims.[21]

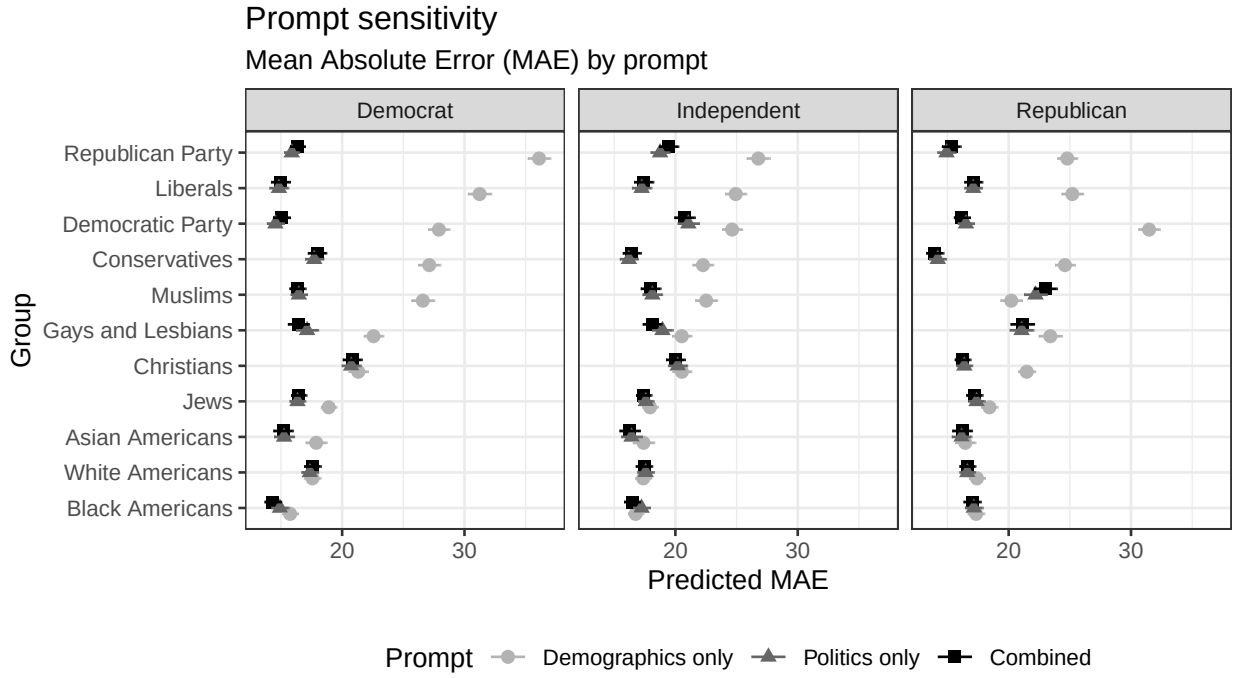

Figure 5: Mean absolute error (MAE, x-axes) associated with different target groups (y-axes) by partisanship (columns) for different prompts to generate the synthetic data. MAE is calculated as the absolute difference between the human respondent's feeling thermometer score for a given target group in the ANES data, relative to the average of 30 synthetic respondents drawn who match the human respondent in terms of their demographics only (light gray circles), political attributes only (dark gray triangles), or both demographic and political attributes combined (black squres).

4.2 Effect of Changes in ChatGPT Over Time

A second dimension of sensitivity concerns time, specifically when the synthetic data are collected. One major question about LLMs is the degree to which they are reproducible. This concern is particularly pertinent to a closed source algorithm like ChatGPT (Spirling 2023). To explore the impact of changes over time we ran a simpler version of our main prompt three times: once in

[21] An additional sensitivity test involves the way in which we use pronouns in our prompt. The synthetic data used in our main analyses relies on a second-person prompt ("You are a ..."). Levendusky and Malhotra (2016) demonstrate that asking humans to imagine being someone else produces exaggerated estimates of polarization. In the SI Section 3, we re-gather synthetic data using a first-person prompt ("I am a ..."), finding less evidence of exaggerated polarization, although worse overall performance in terms of mean absolute error.

April 2023, once in June 2023, and then again in July 2023.[22] In the interim between our June and July runs, OpenAI upgraded its default 3.5 Turbo version of ChatGPT on June 25, 2023, promising that the original will be accessible until September 2023. These three snapshots of the synthetic data allow us to characterize the degree to which such data is "reproducible", which we define in two ways. First, is a researcher able to exactly recover the same dataset over time? Second, is the researcher able to produce a substantively similar dataset, meaning that conclusions drawn from it would persist, even if the cell-by-cell values change? [23] These questions speak to the broader concern over how much the largely opaque decisions of for-profit private companies affect replication.

To test this, we compare our results from the April 2023 data to the June and July 2023 runs of the identical prompt. We plot the April results on the x-axis in Figure 6 and the June and July results on the y-axes. Each point is the count of human-synthetic observations that share a given April-June/July coordinate FT score, aggregating over all target groups. Perfect replication would produce a 45-degree line, which we indicate with a dashed red line.

Neither re-collection of our data – using an identical prompt – exactly reproduces our original synthetic data. However, the July sample is substantially tempered, with the coldest thermometer scores from April increasing and, to a lesser degree, the warmest scores from April declining. Meanwhile the June sample, although it exhibits similar patterns, is not significantly different from a placebo test in which we randomly sample 10 synthetic responses from the April results and treat these as a pseudo-new sample, reflecting mean reversion. The two facets of Figure 6 capture the two sources of variation in synthetic data generated by closed-source LLMs. The left panel highlights that, without the ability of researchers to effectively set a seed for the random number

[22] This prompt was based on our original submission and it did not include descriptions of the respondent's ideology, registration status, or interest in news and politics, and only specified that they were living in the United States in 2019. In addition, we only collected 20 synthetic respondents per persona, instead of 30 per individual human respondent, and we did not record explanations or the confidence in the output. Finally, our June 2023 run of the prompt faced a change in the formatting of the ChatGPT response which meant we did not record any data for one of the target groups. A detailed description of this prompt is included in the SI Section 1.

[23] In investigating these questions, we speak to recent unpublished work which suggests that ChatGPT 4 is getting worse over time, while 3.5 is getting better (Chen et al. 2023), although there is some debate over whether this analysis reveals changes in the LLM's *capabilities* or merely in its *behaviors* (Narayanan and Kapoor 2023).

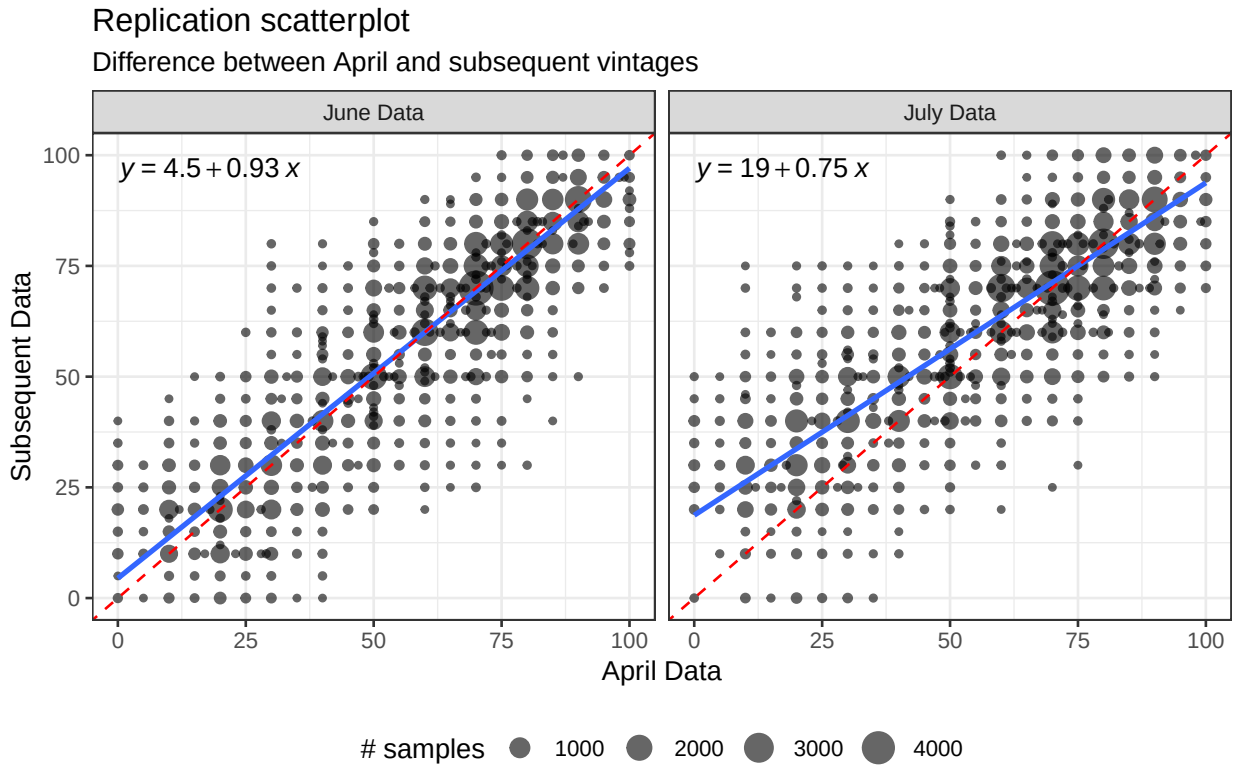

Figure 6: Reproducibility of synthetic data over time. Both plots compare the synthetic dataset generated by a simple prompt gathered in April of 2023 to the identical prompt re-run in June (left facet) and July (right facet) of the same year. Each point indicates the number of observations associated with April vs later synthetic datasets, aggregating across respondents and target groups. Linear regression equation indicated in top-left of both facets, revealing substantially attenuated differences between the April and July runs of the same prompt.

generator, exactly replicating synthetic data is impossible. More concerningly, the right panel highlights the vulnerability of academic norms of replication to the opaque decisions of for-profit private companies. While we still have access to the original version of the ChatGPT 3.5-turbo endpoint that produced our April data at the time of writing, OpenAI has notified its users that this will no longer be available by September of 2023, meaning that whatever conclusions were drawn on this initial sample cannot be reproduced.[24]

More generally, the sensitivity we document along the dimensions of prompt engineering and timing raises concerns for the use of synthetic data for social science research in two ways. First is the issue of reproducibility. While journals might accept the argument that synthetic samples should

[24] We document the original findings of an earlier draft of this manuscript in the Supporting Information Section 6, for posterity.

not be held to higher standards of recreation than human samples, the fact that we find signed differences over time is troubling, especially given that we don't know whether the compression we document is mean reversion or intentional. Second is the issue of prompt sensitivity, which we worry is at the top of a slippery slope of p-hacking if synthetic datasets become acceptable for scholarly research. Given the speed and low costs of creating these data, researchers might be tempted to keep trying different prompts until they assemble the synthetic dataset that generates the results they want. For both of these reasons, we argue that the promise of synthetic data is greatly overstated.

4.3 New and Improved Models?

Despite the evidence presented in this paper, one might hope that there is a future for cheap synthetic data as these models improve. Perhaps the shortcomings we document above are just the growing pains of a sophisticated language model that will one day allow researchers to evaluate social science research questions at a fraction of the cost. We test this claim by comparing the synthetic data generated by ChatGPT 3.5 against the same prompt run using ChatGPT 4.0, finding little support for this optimism.

Figure 7 summarizes the updated AI's performance in terms of basic correlations with the ANES data, relative to version 3.5, revealing no improvement. If anything, ChatGPT 4.0 is slightly worse overall, with a higher y-intercept (suggesting bias away from the most negative feeling thermometer assessments) and a slightly shallower slope (ideally this would be a one-to-one match with perfect synthetic data). These basic patterns are supported in a replication of the analyses above (found in SI Section 11) which demonstrates that ChatGPT 4.0 is a toss-up relative to 3.5 in terms of mean absolute error; continues to generate synthetic samples with smaller deviations from the human sample when predicting feeling thermometers toward ideological groups, but larger deviations when predicting feelings toward parties; and still appears to be overdetermined by partisanship, neglecting more complex conditional relationships, such as age. On net, despite its substantially larger size and demonstrated superior performance across a range of tasks [CITE], our analysis suggests that improvements in ChatGPT's underlying language model have not increased its ability to generate synthetic data.

We sympathize with growing concerns about the reproducibility of research performed with

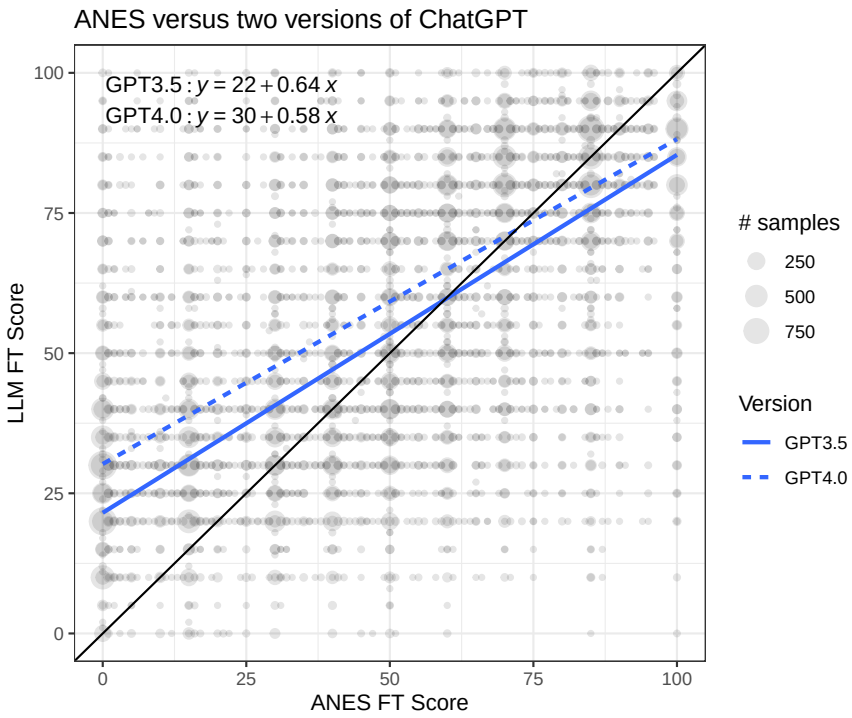

Figure 7: Scatterplot of human feelings toward various outgroups (x-axis) against their associated synthetic samples generated with ChatGPT 3.5-turbo (solid blue line) or ChatGPT 4.0 (dashed blue line). Points sized by the number of respondents associated with each pair of scores.

proprietary models (Spirling 2023), so we also replicated our work using the open source Falcon-40B-Instruct LLM.[25] This was the most powerful open source LLM at the time we began our research,[26] yet we find the same major issues as with ChatGPT. There is considerably less variability in synthetic respondents' feelings toward the two major parties than in the ANES benchmark, and there are numerous (statistically and substantively) significant differences between the partial correlations we would estimate with the Falcon data and those from the survey of real humans.

5 Discussion and Implications

Human subjects are expensive and complicated. If scholars could replace them with LLM-based synthetic subjects, they could much more easily collect information on public opinion, pretest survey questions, substitute the need to actually reach hard-to-reach populations, try out experimental designs, and so on. The allure of closed-source LLMs like ChatGPT is the ability to quickly and

[25] Full details in Supporting Information Section 14.

[26] It has since been overtaken by Meta's Llama-70B.

cheaply obtain data without dealing with the many complications and ethical considerations that are associated with human subject research. The fact that LLMs do a remarkably good job at recovering average responses given by broad groups of respondents in the ANES makes their use all the more tempting. However, we find evidence of several troubling patterns that raise serious concerns about the use of LLMs as a substitute for characterizing public opinion. Not only is precision often excessively high – raising issues with using synthetic responses for power analyses and research design – but the complex interdependencies present among human respondents are not captured well by synthetic responses. This subtle problem can lead to misleading inferences, sometimes even with the wrong sign. In principle an LLM could be developed or trained to perform better on the counts of precision and conditional relationships but to date new versions have not done so.

These results confirm and expand upon the demonstrated limitations that occur when using LLMs to learn about humans. Others have argued that LLMs' understanding of humanity is biased in myriad ways: toward western culture (Washington Post 2023; Cao et al. 2023), toward a progressive sensibility (Motoki et al. 2023; Rozado 2023), and toward a set of personality traits currently reified in the same culture (Rutinowski et al. 2023; Abdulhai et al. 2023; Cao et al. 2023). In line with these results, Section 4 of the appendix confirms that prompting ChatGPT to adopt the persona of an "average" voter or citizen produces responses that are closer to those of Democrats than Republicans. But our contribution is more general in its conclusion that LLMs are unlikely to provide a cheap alternative to human survey data in the near future, and perhaps ever. Although we focus our analysis on ChatGPT, the current best-in-class version of this new technology, we show that open-source versions perform even worse in SI Section 14. Furthermore, our results are based on what we argue is the lowest hanging fruit for the algorithm to generate plausible synthetic samples. When we attempt to replicate our results using the same questions in a national online survey, or using other items in other surveys, or even using other items in surveys conducted in other countries, the performance is even frequently worse than we document here (see SI Sections 12 and 13).[27]

[27] We recognize that treating any sample as a baseline comparison assumes a high level of accuracy about that sample. As discussed in the introduction, the growing crisis in public opinion polling raises questions about whether this assumption is valid. Perhaps the synthetic data are closer to the unknown population parameters of interest, and

Does this mean that the promise of synthetic samples produced by a sophisticated AI will never materialize? We find little evidence that ChatGPT 4.0, at almost 6 times the size of its predecessor, represents a meaningful improvement when it comes to synthetic survey data generation. But perhaps some future version of this technology may one day be able to recover not just basic averages and standard deviations, but also partial associations of interest to researchers. Furthermore, there may be examples of carefully tailored synthetic samples where scholars are able to overcome the limitations we document above, and we can see the value of these applications to impute missing data or to ensure data privacy when making replication materials public. However, without more robust insulation against prompt engineering, which can make the synthetic data draw any number of conclusions, the widespread use of synthetic data as a replacement for human-generated surveys seems unlikely.[28]

The focus of our analyses has been on the practical issue of the quality of synthetic data. We highlight a number of problems with this data source which could lead researchers to draw erroneous conclusions about real human respondents if they were to use it. We want to conclude with a comment about another dimension along which synthetic data from LLMs can be evaluated. The ethics of replacing human opinion with synthetic opinion generated from unknown and unknowable methods seem tenuous at best. Relying on predictions generated by an unknown corpus and using a model with unknown assumptions as a substitute for asking humans how they think and feel about the world around them seems contrary to the origins and importance of polling itself. Polls are intended to check political power and track how opinions change over time and vary between groups. To remove humans from the equation and rely on existing content to extrapolate opinions hard-wires the past into the present. Crucially, it also relies on the voices of the content's creators to characterize the voices of others. Removing or reducing the centrality of humanity from social science and focusing on the black-boxed output of an LLM shifts our attention from characterizing and learning about the opinions and behavior of human beings – however imperfect and complicated those efforts may be – to studying outputs from an algorithm whose relationship to humanity is

the human data is flawed by non-ignorable non-response. However, short of designing and fielding a survey of human respondents that overcomes the challenges that the $14 million ANES cannot, this argument seems untestable and, if true, calls into question decades of public opinion research.

[28]This indictment doesn't consider other dimensions of concern described in Bender et al. (2021) and Spirling (2023).

mostly unknown and, as we show in this study, unable to reproduce the nuances of human beliefs that are of foundational interest to social science.

References

M. Abdulhai, C. Crepy, D. Valter, J. Canny, and N. Jaques. Moral foundations of large language models. 2023.

J. L. Angel and R. A. Settersten Jr. The new realities of aging: Social and economic contexts. *New directions in the sociology of aging*, pages 95–119, 2013.

L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, page 1–15, 2023. doi: 10.1017/pan.2023.2.

C. Bail. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press, 2022.

N. Beauchamp. Predicting and interpolating state-level polling using twitter textual data. *American Journal of Political Science*, 61(2):490–503, 2017.

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

A. J. Berinsky. *Silent voices: Public opinion and political participation in America*. Princeton University Press, 2004.

J. Bisbee. Barp: Improving mister p using bayesian additive regression trees. *American Political Science Review*, 113(4):1060–1065, 2019. doi: 10.1017/S0003055419000480.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2022.

J. Brehm. *The Phantom Respondents: Opinion Surveys and Political Representation*. Michigan Studies In Political Analysis. University of Michigan Press, 1993. ISBN 9780472095230. URL https://books.google.com/books?id=lA1oAAAAIAAJ.

Y. Cao, L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik,

Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.c3nlp-1.7.

D. Caughey and C. Warshaw. Dynamic estimation of latent opinion using a hierarchical group-level irt model. *Political Analysis*, 23(2):197–211, 2015. ISSN 10471987, 14764989. URL http://www.jstor.org/stable/24572968.

A. Cavari and G. Freedman. Survey nonresponse and mass polarization: The consequences of declining contact and cooperation rates. *American Political Science Review*, page 1–8, 2022. doi: 10.1017/S0003055422000399.

L. Chen, M. Zaharia, and J. Zou. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023.

J. D. Clinton, J. J. Agiesta, C. J. Burge, M. Connelly, A. Edwards-Levy, B. Fraga, E. Guskin, D. S. Hillygus, C. Jackson, J. Jones, S. Keeter, K. Khanna, J. Lapinski, L. Saad, D. Shaw, A. E. Smith, M. C. Thee-Brenan, D. Wilson, and C. Wlezien. American Association of Public Opinion Research Task Force on Pre-Election Polling: An evaluation of the 2020 general election polls. https://www.aapor.org/About-Us/Leadership/Committees-and-Taskforces.aspx?cid=2020ELECTIONOnline access, 2021a.

J. D. Clinton, J. S. Lapinski, and M. J. Trussler. Reluctant Republicans, Eager Democrats?: Partisan Nonresponse and the Accuracy of 2020 Presidential Pre-election Telephone Polls. *Public Opinion Quarterly*, 86(2):247–269, 05 2022. ISSN 0033-362X. doi: 10.1093/poq/nfac011. URL https://doi.org/10.1093/poq/nfac011.

I. Cornelis, A. Van Hiel, A. Roets, and M. Kossowska. Age differences in conservatism: Evidence on the mediating effects of personality and cognitive style. *Journal of personality*, 77(1):51–88, 2009.

T. Cowen. Chatgpt ai could make democracy even more messy, Dec 2022. URL https://www.bloomberg.com/opinion/articles/2022-12-06/chatgpt-ai-could-make-democracy-even-more-messy.

M. DeBell and J. A. Krosnick. Computing weights for american national election study survey data. *nes012427. Ann Arbor, MI, Palo Alto, CA: ANES Technical Report Series*, 2009.

J. N. Druckman and M. S. Levendusky. What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1):114–122, 2019.

A. Gelman. Poststratification into many categories using hierarchical logistic regression. *Survey methodology*, 23:127, 1997.

A. S. Gerber, G. A. Huber, D. Doherty, C. M. Dowling, and S. E. Ha. Personality and political attitudes: Relationships across issue domains and political contexts. *American Political Science Review*, 104(1):111–133, 2010.

Y. Ghitza and A. Gelman. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776, 2013.

Y. Ghitza and A. Gelman. Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research. *Political Analysis*, 28

(4):507–531, Oct. 2020. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2020.3. URL http://www.cambridge.org/core/journals/political-analysis/article/voter-registration-databases-and-mrp-toward-the-use-of-largescale-databases-in-public-opinio C6C428EB05DC7132678215896F38B6B7. Publisher: Cambridge University Press.

F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

N. D. Glenn. Aging and conservatism. *The Annals of the American Academy of Political and Social Science*, 415(1):176–186, 1974.

M. Goplerud. Re-evaluating machine learning for mrp given the comparable performance of (deep) hierarchical models. *American Political Science Review*, 2023.

D. A. Graham. The Polling Crisis Is a Catastrophe for American Democracy — theatlantic.com. https://www.theatlantic.com/ideas/archive/2020/11/polling-catastrophe/616986/, 2023. [Accessed 21-Apr-2023].

D. Halpern and J. Gibbs. Social media as a catalyst for online deliberation? exploring the affordances of facebook and youtube for political expression. *Computers in Human Behavior*, 29(3): 1159–1168, 2013.

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

S. Iyengar and S. J. Westwood. The origins and consequences of affective polarization. *Annual Review of Political Science*, 22(1):129–146, 2019.

S. Keeter. The impact of survey non-response on survey accuracy. In *The Palgrave Handbook of Survey Research*, pages 373–381. Springer, 2018.

S. Keeter, N. Hatley, C. Kennedy, and A. Lau. What Low Response Rates Mean for Telephone Surveys. Pew Research Center. https://www.pewresearch.org/methods/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/, May 2017.

C. Kennedy, M. Blumenthal, S. Clement, J. D. Clinton, C. Durand, C. Franklin, K. McGeeney, L. Miringoff, K. Olson, D. Rivers, L. Saad, G. E. Witt, and C. Wlezien. An Evaluation of the 2016 Election Polls in the United States. *Public Opinion Quarterly*, 82(1):1–33, 02 2018. ISSN 0033-362X. doi: 10.1093/poq/nfx047. URL https://doi.org/10.1093/poq/nfx047.

N. Lapidot-Lefler and A. Barak. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2):434–443, 2012.

J. R. Lax and J. H. Phillips. How should we estimate public opinion in the states? *American Journal of Political Science*, 53(1):107–121, 2009.

M. S. Levendusky and N. Malhotra. (mis) perceptions of partisan polarization in the american public. *Public Opinion Quarterly*, 80(S1):378–391, 2016.

J. Mellon and C. Prosser. Correlation with time explains the relationship between survey nonresponse and mass polarization. *The Journal of Politics*, 83(1):390–395, 2021. doi: 10.1086/709433.

B. D. Meyer, W. K. C. Mok, and J. X. Sullivan. Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226, November 2015. doi: 10.1257/jep.29.4.199. URL https://www.aeaweb.org/articles?id=10.1257/jep.29.4.199.

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

A. Narayanan and S. Kapoor. Is gpt-4 getting worse over time? https://www.aisnakeoil.com/p/is-gpt-4-getting-worse-over-time, 2023. Accessed: July 20, 2023.

OpenAI. Chatgpt 3.5 turbo. https://openai.com/blog/chat-gpt-3-5-turbo/, 2021. Accessed: April 22, 2023.

J. C. Peterson, K. B. Smith, and J. R. Hibbing. Do people really become more conservative as they age? *The Journal of Politics*, 82(2):600–611, 2020.

P. Rossini. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425, 2022.

I. Rowe. Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, communication & society*, 18(2):121–138, 2015.

D. Rozado. The political biases of chatgpt. *Social Sciences*, 12(3), 2023. ISSN 2076-0760. doi: 10.3390/socsci12030148. URL https://www.mdpi.com/2076-0760/12/3/148.

J. Rutinowski, S. Franke, J. Endendyk, I. Dormuth, and M. Pauly. The self-perception and political biases of chatgpt. *arXiv preprint arXiv:2304.07333*, 2023.

S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect?, 2023.

W. Shapiro. The polling industry is in crisis, June 21 2019. URL https://newrepublic.com/article/154124/polling-industry-crisis.

A. Spirling. Why open-source generative ai models are an ethical way forward for science. *Nature*, 616:413, 2023. doi: https://doi.org/10.1038/d41586-023-01295-4.

P. Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.

J. Tucker. Measuring public opinion with social media data. In *The Oxford Handbook of Polling and Polling Methods*, pages 1–22. Oxford University Press, 2017.

M. van Klingeren, D. Trilling, and J. Möller. Public opinion on twitter? how vote choice and arguments on twitter comply with patterns in survey data, evidence from the 2016 ukraine referendum in the netherlands. *Acta Politica*, 56(3):436–455, 2021. doi: 10.1057/s41269-020-00160-w. URL https://doi.org/10.1057/s41269-020-00160-w.

D. Waldner and E. Lust. Unwelcome change: Coming to terms with democratic backsliding. *Annual Review of Political Science*, 21:93–113, 2018.

W. Wang, D. Rothschild, S. Goel, and A. Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015. Publisher: Elsevier.

T. Washington Post. Ai chatbots are learning to hold more natural conversations. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/, 2023. Accessed: April 22, 2023.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models, 2022.

P. Y. Wu, J. A. Tucker, J. Nagler, and S. Messing. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057*, 2023.

Supporting Information for Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

James Bisbee* Joshua D. Clinton† Cassy Dorff‡

Brenton Kenkel§ Jennifer Larson¶

August 9, 2023

Contents

*Assistant Professor of Political Science, Vanderbilt University james.h.bisbee@vanderbilt.edu. Corresponding author.
†Abby and Jon Wikelried Professor, Vanderbilt University josh.clinton@vanderbilt.edu
‡Assistant Professor of Political Science, Vanderbilt University cassy.dorff@vanderbilt.edu
§Assistant Professor of Political Science, Vanderbilt University brenton.kenkel@vanderbilt.edu
¶Associate Professor of Political Science, Vanderbilt University jennifer.larson@vanderbilt.edu

1 Prompt Engineering

Creating the final prompt used for the main results was an iterative process. We initially tried simply asking the LLM to imagine that they were a certain persona and to answer questions. This produced results that were both difficult to process (open-ended text results) as well as often resulting in the LLM's refusal to provide the answers requested. We found that using the instruction "Provide responses from this person's perspective. Use only knowledge about politics that they would have." helped avoid refusals by the AI, although the resulting output remained open-ended and hard to process. We then provided more detailed instructions about the format of the output requested, describing the the desired .csv format where each row corresponded to a given target group, and the columns included the name of the group and the synthetic thermometer score.

The code that produced the resulting prompt for our initial synthetic data was used in early April of 2023, and is reproduced below. We relied on the `openai` package for `R` and described the persona along the dimensions of age, race, gender, income (`inc`), education (`educ`), and party ID (`pid`). All persona were defined as registered voters who lived in the United States in 2019. The following approach was inspired by Marquez [2023].

```
library(tidyverse)
library(openai)

Sys.setenv(OPENAI_API_KEY = 'YOUR_KEY_HERE')

# Function to create prompt out of inputs
create_prompt <- function(audit_data) {
  res <- list()
  for(i in 1:nrow(audit_data)) {
    age = audit_data$age[i]
    race = audit_data$race[i]
    gender = audit_data$gender[i]
    inc = audit_data$inc[i]
    educ = audit_data$educ[i]
    pid = audit_data$pid[i]
    res[[i]] <- list(
      list(
        "role" = "system",
        "content" = stringr::str_c(
          "You are a ",age," year old ",race," ",gender,
          " with a ",educ,", earning $",inc," per year. ",
          "You are a registered ",pid," living in the USA in 2019.")
      ),
      list(
        "role" = "user",
        "content" = stringr::str_c(
          "Provide responses from this person's perspective.\n
          Use only knowledge about politics that they would have.\n
          Format the output as a csv table with the following format:\n
          group,thermometer\n
          The following questions ask about individuals' feelings
          toward different groups.\n
          Responses should be given on a scale from 0 (meaning cold
          feelings) to 100 (meaning warm feelings).\n
          Ratings between 50 degrees and 100 degrees mean that\n
          you feel favorable and warm toward the group. Ratings
```

```
          between 0\n
          degrees and 50 degrees mean that you don't feel
          favorable toward\n
          the group and that you don't care too much for that
          group. You\n
          would rate the group at the 50 degree mark if you don't feel\n
          particularly warm or cold toward the group.\n
          How do you feel toward the following groups?\n",
          'The Democratic Party?\n',
          'The Republican Party?\n',
          'Democrats?\n',
          'Republicans?\n',
          'Black Americans?\n',
          'White Americans?\n',
          'Hispanic Americans?\n',
          'Asian Americans?\n',
          'Muslims?\n',
          'Christians?\n',
          'Immigrants?\n',
          'Gays and Lesbians?\n',
          'Jews?\n',
          'Liberals?\n',
          'Conservatives?\n',
          'Women?\n')
      )
    )
  }

  return(res)
}

# Define profiles to iterate over
audit_data <- expand.grid(age = c(20,35,50,65),
                          race = c('non-Hispanic white',
                                   'non-Hispanic black',
                                   'Hispanic'),
                          gender = c('male','female'),
                          inc = c('30,000','50,000','80,000',
                                  '100,000','more than $150,000'),
                          educ = c('high school diploma',
                                   "some college, but no degree",
                                   "bachelor's degree",
                                   "postgraduate degree"),
                          pid = c('Republican','Democrat','Independent'),
                          stringsAsFactors = F) %>%
  as_tibble()

# Function to submit the query
submit_openai <- function(prompt, temperature = 0.2, n = 1) {
  res <- openai::create_chat_completion(model = "gpt-3.5-turbo",
                                        messages = prompt,
                                        temperature = temperature,
                                        n = n)
  Sys.sleep(1)
```

```
  res
}

# Create an empty csv file to append to.
# df <- data.frame(audit_data[0,],
#                  group = as.character(),
#                  thermometer = as.numeric(),
#                  draw = as.numeric(),
#                  index = as.numeric(),
#                  stringsAsFactors = F)
#
# write.table(df,file = './results/therm_ANES.csv',
#             append = F,row.names = F,col.names = T,sep = ',')

# Load the already completed data
df <- read_csv('./results/therm_ANES.csv') %>%
  mutate(index = as.numeric(gsub(',NA','',index)))

# Pick up where the previous run left off
if(nrow(df) == 0) {
  start = 1
} else {
  start = max(df$index,na.rm=T) + 1
}

toSave <- NULL
TPM <- RPM <- NULL
zz <- zzz <- Sys.time()
for(i in start:nrow(audit_data)) {
  prompts <- create_prompt(audit_data[i,])

  # Iterate over different temperature settings
  for(t in seq(.1,1,by = .3)) {
    openai_completions <- try(prompts |>
                              purrr::map(submit_openai,temperature = t,n =
    20))

    while(class(openai_completions) == 'try-error') {
      Sys.sleep(60)
      cat('issue on\n',
          'temp =',t,'\n',
          paste(audit_data[i,],collapse = ' / '),'\n')
      openai_completions <- try(prompts |>
                                purrr::map(submit_openai,temperature = t,n
    = 20))

    }

    tmp <- NULL
    for(j in 1:length(openai_completions[[1]]$choices$message.content)) {
      tmp <- bind_rows(tmp,
                       read.csv(text = gsub('\\\\','',
                        openai_completions[[1]]$choices$message.content[j]
   ),
```

```
                                    col.names = c('group','thermometer')) %>%
        mutate(draw = j,
                temp = t,
               thermometer = as.numeric(thermometer)))
    }

    toSave <- toSave %>%
      as_tibble() %>%
      bind_rows(data.frame(audit_data[i,]) %>%
                  cbind(tmp %>%
                          mutate(index = i)))

    TPM <- sum(TPM,openai_completions[[1]]$usage$total_tokens)
    RPM <- sum(RPM,1)
  }

  # Code to prevent exceeding API limits
  if(difftime(Sys.time(),zzz,units = 'mins') < 1) {
    if(RPM > 3000 | TPM > 85000) {
      cat('RPM = ',RPM,'\nTPM = ',TPM,'\n')
      Sys.sleep(max(0,as.numeric(60 - difftime(Sys.time(),zzz,units = '
   secs'))))
      RPM <- TPM <- NULL
      zzz <- Sys.time()
      cat('Approaching rate limit\n')
    }
  } else {
    RPM <- TPM <- NULL
    zzz <- Sys.time()
  }

  # Append results to csv file every hundred profiles
  if(i %% 100 == 0) {
    write.table(toSave,file = './results/therm_ANES.csv',
                append = T,row.names = F,col.names = F,sep = ',')
    toSave <- NULL

    cat(i,'in',round(difftime(Sys.time(),zz,units = 'mins'),2),'minutes\n'
   )
    zz <- Sys.time()
  }
}
```

We ran the preceding prompt three times in the spring and summer of 2023: first in mid-April, second in mid-June, and again in early July. In each case, there were rare instances in which either the code or the API experienced an issue that prevented us from gathering the data for a given target group-by-persona. In most cases, these were random with respect to our quantities of interest. In one case though, there were systematically more issues with our synthetic responses for the Democratic Party. This was due to the fact that we asked about this group first in the list of target groups, meaning it would occupy the first row in the resulting csv/tsv file.[1] In some cases,

[1] We initially requested the results in .csv format, but found that this would create issues with the detailed version of the prompt that asked for explanations, since these would often include commas. As such, we switched to the tsv

this first row would not include a header, meaning that we treated the response for the Democratic Party target group as the header and lost it. These issues were rare enough in the initial April run of the script that we did not notice the issue until later. However, the June iteration of the same prompt yielded many errors of this type, but – perplexingly – only for the prompts for Democrat and Independent personas. The Republican persona was far more likely to include the header row and avoid the issue. We fixed the issue in subsequent runs of the API, and drop the Democratic Party results from all analyses of the June vintage of our data. We plot the API errors below in Figure 1 by party, target group, and vintage, highlighting that, with the exception of the aforementioned issue with the Democratic Party target group, there is little evidence to make us concerned that errors in data collection are non-randomly associated with the target groups or persona profiles. Even in cases where we observe larger proportions missing, these never exceed 1.15% of the intended synthetic sample size. In most cases, missingness amounts to a single synthetic sample out of the 20 intended to be gathered for a given persona for a given target group.

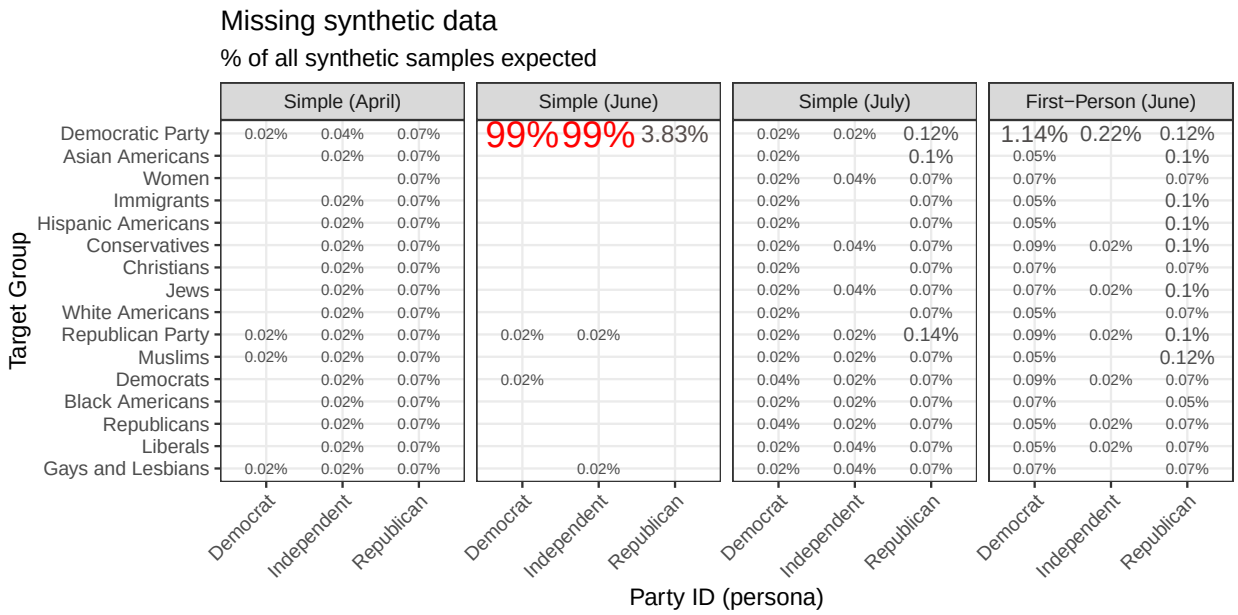

Figure 1: Proportion of missing synthetic responses by target group (y-axes), synthetic party ID (x-axes) and vintage of data collection / prompt type (columns).

In June, we also re-ran the preceding prompt replacing second-person pronouns with first-person pronouns at the suggestion of a reviewer who pointed out that human respondents believe out-groups hold more extreme views than they actually do [Levendusky and Malhotra, 2016]. It is possible therefore that ChatGPT inherits this second-person bias, exaggerating extremism not via any particular limitation of the technology, but rather by accurately reflecting how humans perceive polarization. The revised first-person prompt is copied below (note that the only difference between the original prompt and the first person is found in the `system` parameter, and we save space by only copying the adjusted language).

```
    list(
      list(
        "role" = "system",
```

format in June.

```
      "content" = stringr::str_c(
        "I am a ",age," year old ",race," ",gender,
        " with a ",educ,", earning $",inc," per year. ",
        "I am a registered ",pid," living in the USA in 2019.")
    ),
    list(
      "role" = "user",
      "content" = stringr::str_c(
        "Provide responses to the following questions.\n
        Format the output as a csv table with the following format:\n
        group,thermometer\n
        The following questions ask about individuals' feelings toward
different groups.\n
       Responses should be given on a scale from 0 (meaning cold
feelings) to 100 (meaning warm feelings).\n
       Ratings between 50 degrees and 100 degrees mean that\n
       you feel favorable and warm toward the group. Ratings between 0\
n
       degrees and 50 degrees mean that you don't feel favorable toward
\n
       the group and that you don't care too much for that group. You\n
       would rate the group at the 50 degree mark if you don't feel\n
       particularly warm or cold toward the group.\n
       How do you feel toward the following groups?\n",
       'The Democratic Party?\n',
       'The Republican Party?\n',
       'Democrats?\n',
       'Republicans?\n',
       'Black Americans?\n',
       'White Americans?\n',
       'Hispanic Americans?\n',
       'Asian Americans?\n',
       'Muslims?\n',
       'Christians?\n',
       'Immigrants?\n',
       'Gays and Lesbians?\n',
       'Jews?\n',
       'Liberals?\n',
       'Conservatives?\n',
       'Women?\n')
    )
```

We also created a new prompt that allowed us to more carefully test which aspects of the persona matter most to accuracy of the synthetic data. This prompt operated slightly differently from the preceding, although the core components remained. Specifically, we used this prompt to collect 30 synthetic respondents per human, not per persona. In addition, we also asked the AI to provide both an explanation for its response, as well as a numeric measure of its confidence in its response.

The inclusion of an explanation, although significantly more expensive, was based on an anonymous reviewer's observation that restricting the LLM to provide only numbers may hurt its performance, since numbers live in a similar part of the embedding space. The inclusion of a confidence measure was also based on reviewer feedback, who wondered whether we could actually measure the posterior probability of the thermometer score chosen. We analyze both of these additional

pieces of information in more detail below. In SI Section 5 below, we use NLP on the open-ended explanations provided by ChatGPT to validate the numeric thermometer scores that were provided by showing that the explanations for two responses that are closer together numerically, are also more similar in terms of their language. And in SI Section 9.3 below, we investigate the relationship between the empirical uncertainty across 30 synthetic respondents for a given human, and the "self-reported" confidence of the LLM when it generated those numbers.

The prompt for the open-ended response is reproduced below. (Note that the code has been refined for greater legibility. We now generate the `system` prompt with a simple `gsub()` command that flies in the relevant characteristics directly from the ANES respondent.)

```
full <- "It is [YEAR]. You are a [AGE] year-old, [MARST], [RACETH] [GENDER] with [EDUCATION] making [INCOME] per year, living in the United States.
You are [IDEO], [REGIS] [PID] who [INTEREST] pays attention to what's going on in government and politics."

demog <- "It is [YEAR]. You are a [AGE] year-old [MARST] [RACETH] [GENDER] with [EDUCATION] making [INCOME] per year, living in the United States."

pol <- "It is [YEAR]. You are [IDEO], [REGIS] [PID] who [INTEREST] pays attention to what's going on in government and politics, living in the United States."


system_prep <- function(frame,lookup) {
  for(col in colnames(lookup)) {
    frame <- gsub(paste0('\\[',toupper(col),'\\]'),lookup[[col]][1],frame)
  }
  return(frame)
}


create_prompt <- function(systemPrompt) {
  res <- list(
      list(
        "role" = "system",
        "content" = systemPrompt
      ),
      list(
        "role" = "user",
        "content" = stringr::str_c(
          "Provide responses from this person's perspective.\n
          Use only knowledge about politics that they would have.\n
          Format the output as a tsv table with the following format:\n
          group\tthermometer\texplanation\tconfidence\n
          The following questions ask about individuals' feelings toward different groups.\n
          Responses should be given on a scale from 0 (meaning cold feelings) to 100 (meaning warm feelings).\n
          Ratings between 50 degrees and 100 degrees mean that\n
          you feel favorable and warm toward the group. Ratings between 0\n
```

```
          degrees and 50 degrees mean that you don't feel favorable toward \n
          the group and that you don't care too much for that group. You\n
          would rate the group at the 50 degree mark if you don't feel\n
          particularly warm or cold toward the group.\n
          How do you feel toward the following groups?\n",
          'The Democratic Party?\n',
          'The Republican Party?\n',
          'Democrats?\n',
          'Republicans?\n',
          'Black Americans?\n',
          'White Americans?\n',
          'Hispanic Americans?\n',
          'Asian Americans?\n',
          'Muslims?\n',
          'Christians?\n',
          'Immigrants?\n',
          'Gays and Lesbians?\n',
          'Jews?\n',
          'Liberals?\n',
          'Conservatives?\n',
          'Women?\n')
      )
    )
  return(res)
}
```

We ran this detailed prompt over the course of several weeks in June, 2023, using ChatGPT 3.5 turbo. In addition, we re-ran a smaller version of the prompt in July of 2023, using ChatGPT 4.0. To reduce expenses associated with the more costly GPT 4.0 endpoint, we only collected one synthetic response per human, and only queried four target groups: Democrats, Republicans, Liberals, and Conservatives. Data missingness in the detailed prompt using ChatGPT 3.5 again does not appear to 3.5 be correlated with a particular group or persona, with the exception of the Democratic (and, to a lesser extent, Republican) party (see Figure 2). Missingness in the GPT 4 version amounts to 11 missing human respondents without a synthetic counterpart at all, and – among those with a synthetic counterpart – no missing responses for the Republican Party target group, one missing response for the liberal group, six missing responses for the conservative group, and 22 missing responses for the Democratic Party target group.

Finally, we used a modified version of the simpler prompt to evaluate the performance of ChatGPT 3.5 on the 2016 ISSP dataset. Here, we only describe the human persona in terms of their age, marital status, gender, education, and the country in which they lived. As above, we instructed the LLM to adopt the persona's perspective, provide only the numeric response for each question, and described the Likert scale as it appeared in the original survey. The code to generate synthetic responses to the questions about government are given below:

```
full <- "It is 2016. You are a [AGE] year-old, [MARST], [GENDER] with [
    EDUC], living in [COUNTRY]."

system_prep <- function(frame,lookup) {
  for(col in colnames(lookup)) {
    frame <- gsub(paste0('\\[',toupper(col),'\\]'),lookup[[col]][1],frame)
  }
  return(frame)
```

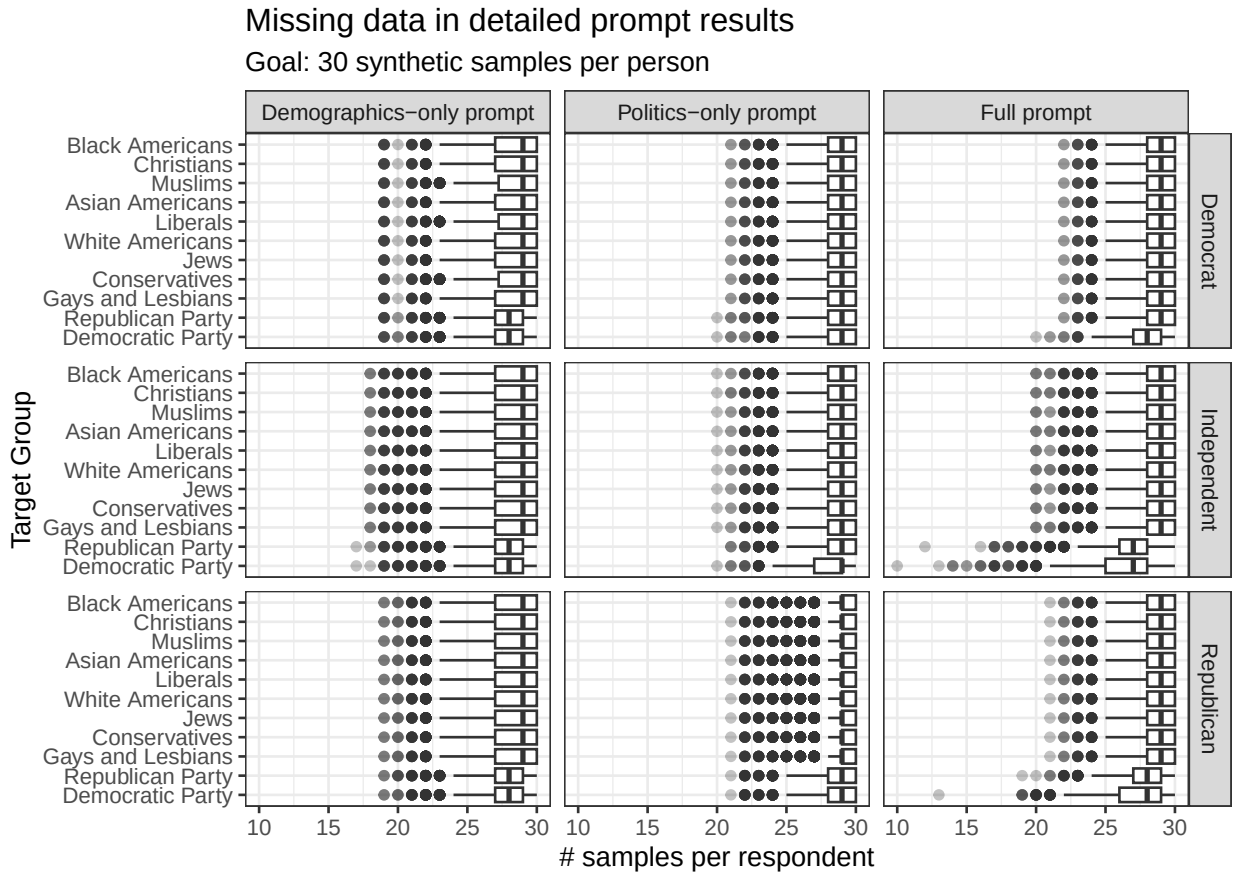

Figure 2: Number of synthetic responses by ANES respondent (x-axes) by target group (y-axes), detailed prompt type (columns), and synthetic party ID (rows).

```
}



submit_openai <- function(prompt, temperature = 0.2, n = 1) {
  res <- openai::create_chat_completion(model = "gpt-3.5-turbo",
                                        messages = prompt,
                                        temperature = temperature,
                                        n = n)
  Sys.sleep(1)
  res
}

create_prompt <- function(systemPrompt) {
  res <- list(
    list(
      "role" = "system",
      "content" = systemPrompt
    ),
    list(
      "role" = "user",
```

```
      "content" = stringr::str_c(
        "Provide responses from this person's perspective.\n
          Use only knowledge about politics that they would have.\n
          Format the output as a tab-separated value (tsv) table with the following format:\n
          question\tanswer\n
          where the question is formatted as Q1, Q2, etc. and the response is formatted numerically.\n
          For example, the first row of the table should be as follows:
          Q1\t1\n
          Q2\t4\n
          Q3\t3\n
        The following questions ask about individuals' feelings regarding the government and its proper role in society.\n
          Each question can be answered on a 5-item scale with the following categories:\n
          1 - Strongly in favour of\n
          2 - In favour of\n
          3 - Neither in favour nor against\n
          4 - Against\n
          5 - Strongly against\n
        'Here are something things the government might do for the economy.
        Please indicate which actions you are in favour of and which you are against.\n
        Q1: Cuts in government spending.\n
        Q2: Government financing of projects to create new jobs.\n
        Q3: Less government regulation of business.\n
        Q4: Support for industry to develop new products and technology.\n
        Q5: Support for declining industries to protect jobs.\n
        Q6: Reduce the working week to create more jobs.\n")
    )
  )
  return(res)
}
```

The code to generate synthetic responses to the questions about protestors and revolutionaries are given below:

```
full <- "It is 2016. You are a [AGE] year-old, [MARST], [GENDER] with [EDUC], living in [COUNTRY]."

system_prep <- function(frame,lookup) {
  for(col in colnames(lookup)) {
    frame <- gsub(paste0('\\[',toupper(col),'\\]'),lookup[[col]][1],frame)
  }
  return(frame)
}



submit_openai <- function(prompt, temperature = 0.2, n = 1) {
  res <- openai::create_chat_completion(model = "gpt-3.5-turbo",
                                        messages = prompt,
                                        temperature = temperature,
```

```
                                          n = n)
  Sys.sleep(1)
  res
}

create_prompt <- function(systemPrompt) {
  res <- list(
    list(
      "role" = "system",
      "content" = systemPrompt
    ),
    list(
      "role" = "user",
      "content" = stringr::str_c(
        "Provide responses from this person's perspective.\n
          Use only knowledge about politics that they would have.\n
          Format the output as a tab-separated value (tsv) table with the following format:\n
          question\tanswer\n
          where the question is formatted as Q1, Q2, etc. and the response is formatted numerically.\n
          For example, the first row of the table should be as follows:
          Q1\t1\n
          Q2\t4\n
          Q3\t3\n
        The following questions ask about individuals' views about how to protest against a government action.\n
          Each question can be answered on a 4-item scale with the following categories:\n
          1 - Definitely allowed\n
          2 - Probably allowed\n
          3 - Probably not allowed\n
          4 - Definitely not allowed\n
        'There are many ways people or organisations can protest against a government action they strongly oppose.\n
        Please show which you think should be allowed and which should not be allowed.\n
        Q1: Organising public meetings to protest against the government.\n
        Q2: Organising protest marches and demonstrations.\n
        There are some people whose views are considered extreme by the majority. Consider people who want to overthrow the government by revolution.\n
        Q3: Do you think such people should be allowed to hold public meetings to express their views?.\n
        Q4: Do you think such people should be allowed to publish books expressing their views?'\n")
    )
  )
  return(res)
}
```

For ease of reference, we refer to the different prompts used in our analysis according to the naming conventions described in Table 1.

Prompt	Model	Description	Response	Date
Simple	GPT 3.5	age, race, gender, education, income, partisanship	20 thermometer scores per 16 target groups-by-1,440 profiles	Apr Jun Jul '23
First-person	GPT 3.5	age, race, gender, education, income, partisanship	20 thermometer scores per 16 target groups-by-1,440 profiles	June 2023
Detailed 3.5	GPT 3.5	age, race, gender, education, income, partisanship, ideology, registration status, news interest, marital status, year	30 thermometer scores, explanations, and confidences per 16 target groups-by-7,530 humans	June 2023
Detailed 4.0	GPT 4.0	age, race, gender, education, income, partisanship, ideology, registration status, news interest, marital status, year	1 thermometer score, explanation, and confidence per 4 target groups-by-7,530 humans	July 2023
ISSP Gov	GPT 3.5	age, marital status, gender, education, country	30 Likert responses per 6 questions-by-2,481 profiles	July 2023
ISSP Prot	GPT 3.5	age, marital status, gender, education, country	30 Likert responses per 4 questions-by-2,481 profiles	July 2023

Table 1: Description of all prompts used. Main results are based on the "Detailed 3.5" prompt, where we calculated the synthetic data's per-human response using the average of its 30 draws.

There is little evidence of our prompts failing to generate estimates in a systematic manner, beyond the issues produced by our code.

2 Temperature and "creativity"

In the context of ChatGPT, "temperature" refers to how deterministic the choice of the subsequent token in its output is, conditional on the prompt and the preceding tokens. Each subsequent token is chosen from a probability distribution across tokens. A temperature setting of zero means that the AI will always choose the most likely subsequent token. As temperature rises, the selection from this distribution grows less deterministic, which is often colloquially referred to as "creativity".

Our main results are generated by ChatGPT at its most "creative", meaning that the temperature hyperparameter was set to 1.[2] In theory, these results should be the noisiest and, potentially, the most representative of the actual randomness in human survey responses. As we demonstrated in the paper, even at this temperature setting, the ChatGPT estimates were far more precise than those found among ANES respondents. On average, LLM estimates were only 40% as variable as their ANES counterparts.

Note that this number combines two sources of variation. The first is the variation stemming from averaging across different groups. For example, the LLM's standard deviations for the party-by-race results incorporated variation stemming from other covariates such as age, gender, educational attainment, and income. The second is the inherent randomness of the data generating process. Among humans, this is a reflection of all our quirks that aren't captured by covariates. In the LLM, it is a characteristic of the model, which can be partially tweaked by the temperature parameter.

With this in mind, how much worse does this overconfidence grow if we reduce the creativity? To investigate, we calculated the standard deviation for each target group for each profile in the LLM data by temperature settings ranging from 0.1 to 1, using the original "simple" prompt from April. We plot the averages of these measures of variance in Figure 3, illustrating how much less uncertain our measures would have been had we reduced the temperature parameter. In all cases, we highlight that reducing the temperature value (x-axes) reduces the average standard deviation (y-axes) across target groups (rows), regardless of how coarse or how granular our aggregation of the personas is (columns).

[2] Technically, the temperature hyperparameter can go even higher than 1. However, in testing we found that exceeding the value of 1 produced either gibberish results, or results that were no longer formatted as requested to aid data extraction.

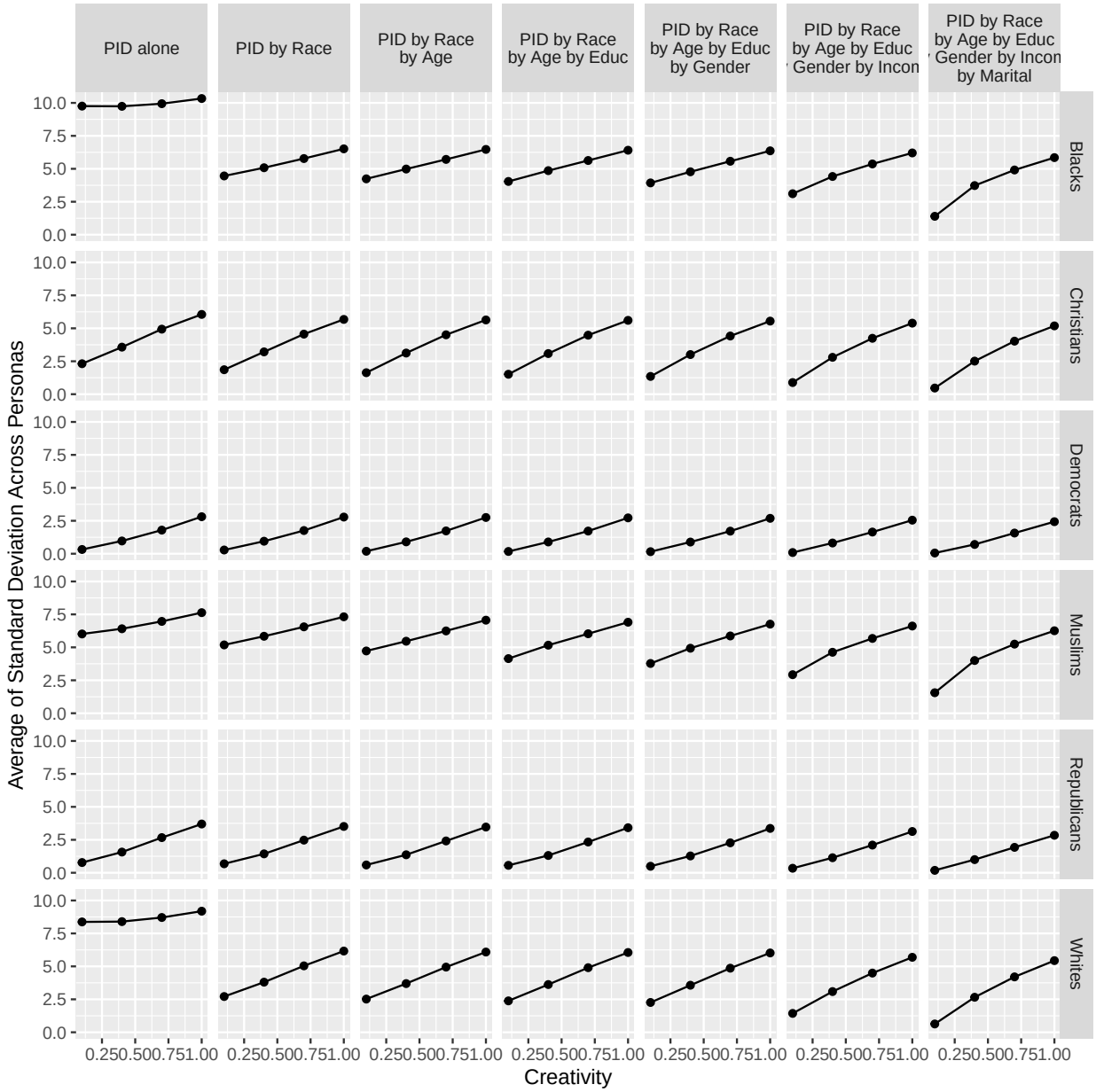

Figure 3: Relationship between temperature hyperparameter (x-axes) and average standard deviation of attitudes (y-axes) by target group (rows) and level of persona aggregation (columns).

3 First-person and second-person prompts

Our main results find that ChatGPT exaggerates polarization in the United States in 2019, compared to actual human survey respondents to the 2016 and 2020 ANES waves. However, existing work demonstrates that human respondents believe out-groups hold more extreme views than they actually do [Levendusky and Malhotra, 2016]. It is possible therefore that ChatGPT inherits this second-person bias, exaggerating extremism not via any particular limitation of the technology, but rather by accurately reflecting how humans *perceive* polarization. As such, we re-ran our prompt on July 5th, 2023 where we instructed ChatGPT with first-person descriptions (i.e. "I am a 24 year old white male with a college degree earning $50,000 per year." See the full prompt description in SI Section 1 above.).

Consistent with expectations, we find attenuated evidence of exaggerated affective polarization when using the first-person prompt (see Figure 4). In particular, we no longer observe more negative attitudes toward out-group ideologues. In addition, at least among synthetic Democrats, the first-person prompt yields greater accuracy in three of the four comparisons to the ANES estimate (the exception being synthetic Democrats' attitudes toward the Republican Party). However, the first-person prompt is *less* accurate in almost every other comparison.

Although the evidence of exaggerated polarization is attenuated with the first person prompt, this does not mean that a first person prompt performs better. As illustrated in Figure 5, the mean absolute error for the first person prompt is larger than the original prompt for all target groups except liberals and Muslims, and this pattern is even worse for the first person prompt when we compare it to the July run of the original prompt.

A final note on the question of first- versus second-person pronouns is that the vast majority of explanations that were recorded with the detailed prompt – which used second-person pronouns to describe the persona to the LLM – were expressed in the first-person. In other words, even though ChatGPT was prompted with "You are a...", it replied with "I", suggesting that it is able to better inhabit the persona suggested than the humans documented in Levendusky and Malhotra [2016], who were asked to indicate what they thought a "typical Democrat voter would want" (i.e., a third-person prompt).

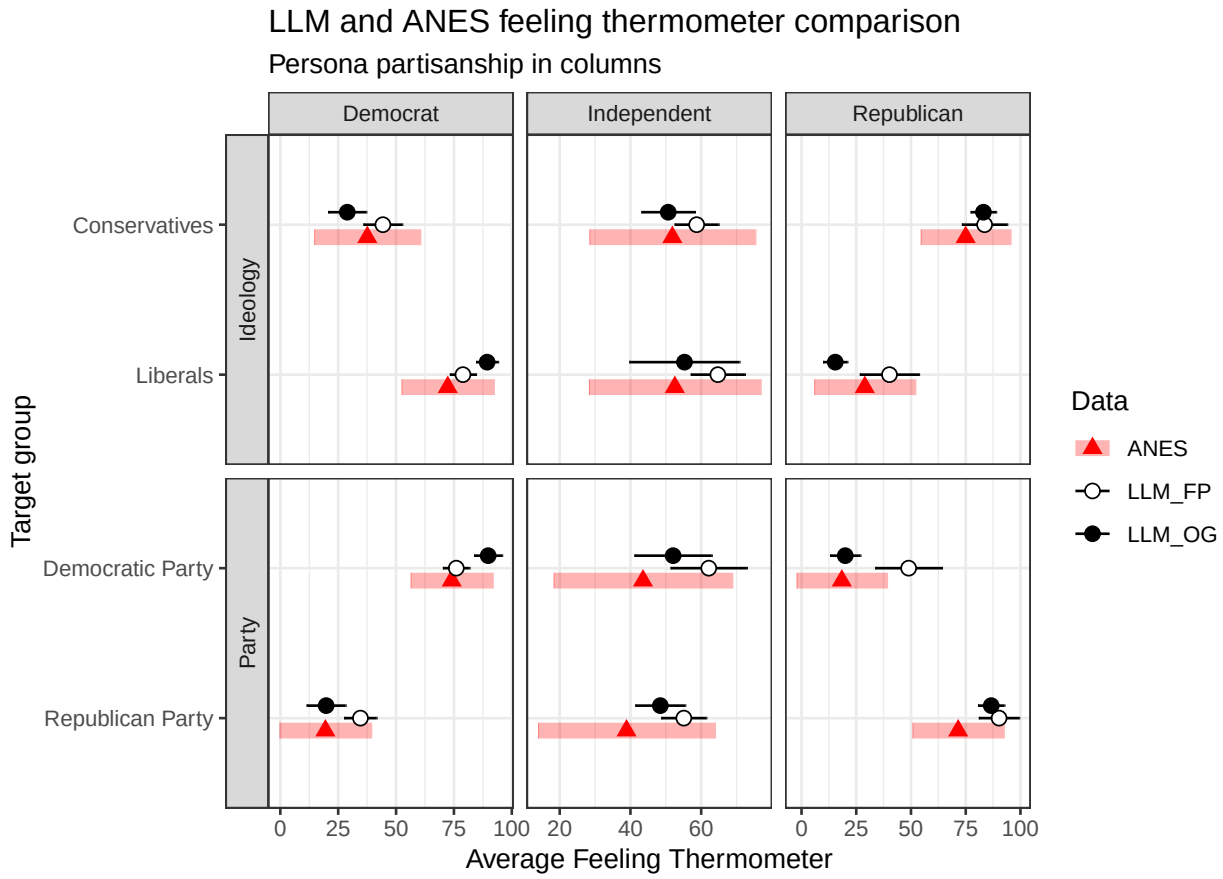

Figure 4: Average feeling thermometer (x-axes) toward political target groups (y-axes) by synthetic party ID (columns) and data type (red triangles are human respondents measured in the ANES, solid black points are synthetic responses measured using the simple prompt in April 2023, and hollow points are synthetic responses measured using the first-person variant of the simple prompt, measured in June).

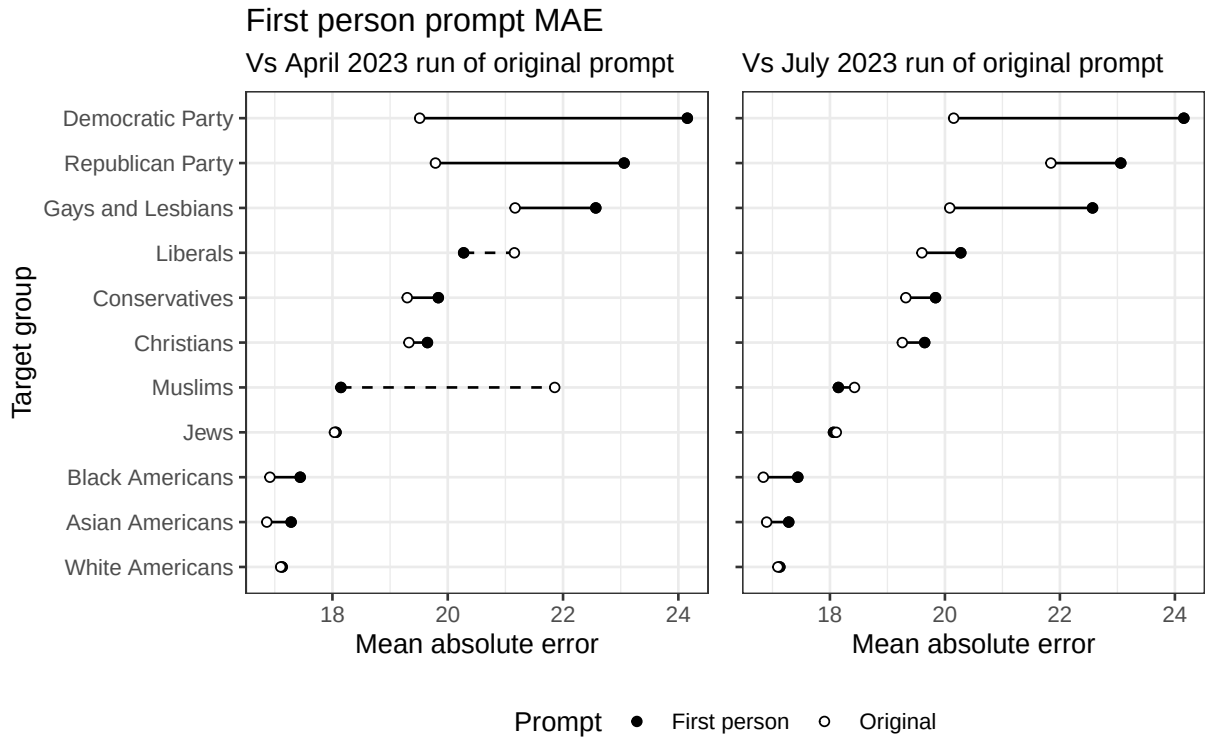

Figure 5: Mean absolute error (x-axes) measuring the average absolute difference between human respondents in the ANES and their synthetic counterparts produced by either the first person version of the simple prompt (black circles) or the second-person version of the simple prompt (hollow circles) by target group (y-axes). Columns indicate whether the second-person version of the simple prompt was collected in April or July of 2023.

4 Replication

We document evidence of variation over time in our manuscript, suggesting that the July run of the simple prompt produced less extreme measures compared to the initial April version of the data. Here, we investigate this comparison in more detail, starting with a summary measure of the change in thermometer scores. Aggregating over all profiles and all target groups, we find that the July 2023 vintage data is less negative overall compared to the April 2023 vintage data. Figure 6 displays the histogram of the difference between the April and July versions of the data, where negative values indicate that the July values are larger than their corresponding April values. As illustrated, 48% of observations were higher in July than in April, compared to only 27% being higher in April. Furthermore the magnitude of these differences is quite large, ranging up to 50 feeling thermometer points on the scale from 0 to 100.

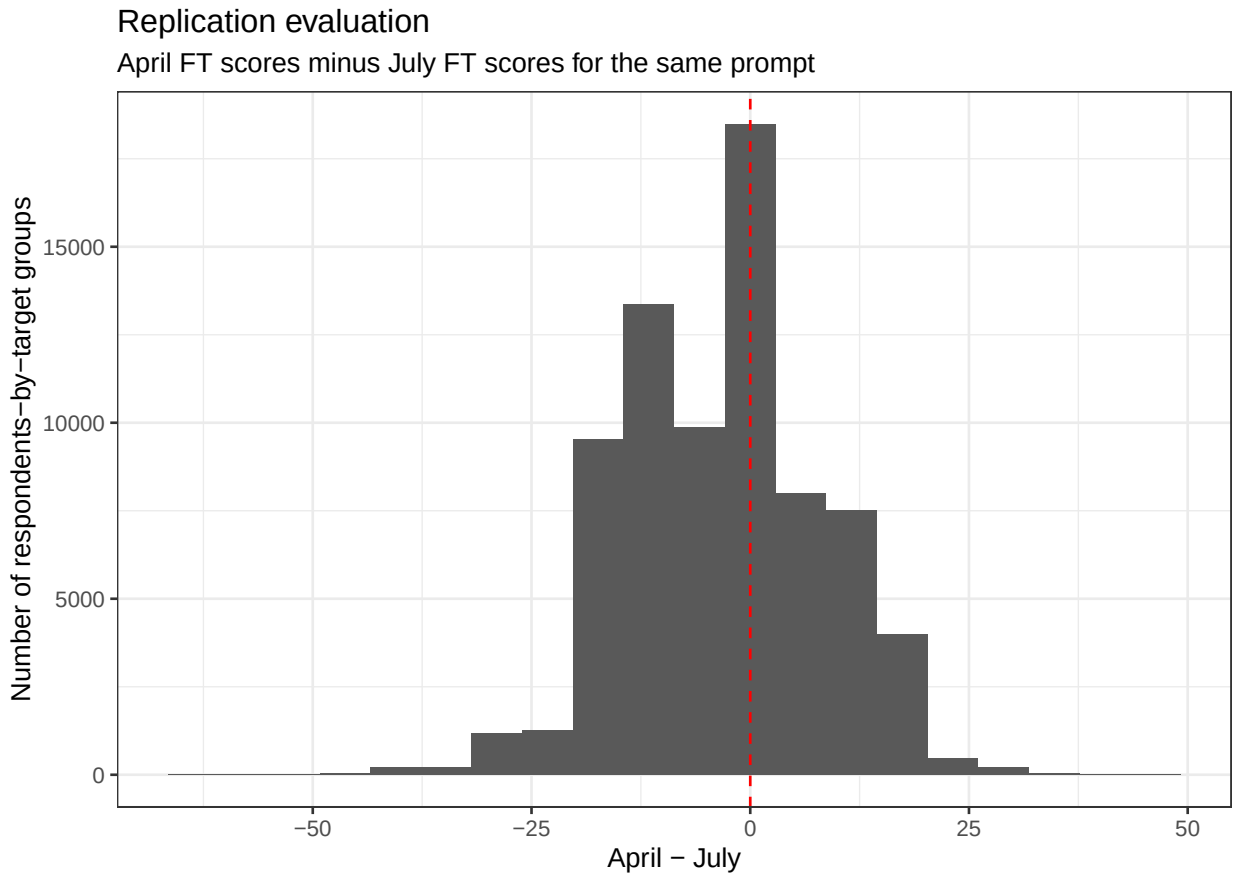

Figure 6: Difference in synthetic estimates of feeling thermometers between the April and July 2023 vintages of the simple prompt.

In our main results, we show that this increase in warmth is driven primarily by the coolest thermometer scores recorded in the April vintage data. Below, we calculate these scatter plots for all target groups, illustrating that the big picture compression is not driven by any particular target group (see Figure 7).

However, even here these patterns mask how weak the association is once we remove party. We reproduce the same plot once more but calculate the associations for Democrats, Independents, and Republicans separately. As illustrated in Figure 8, there is essentially zero correlation between

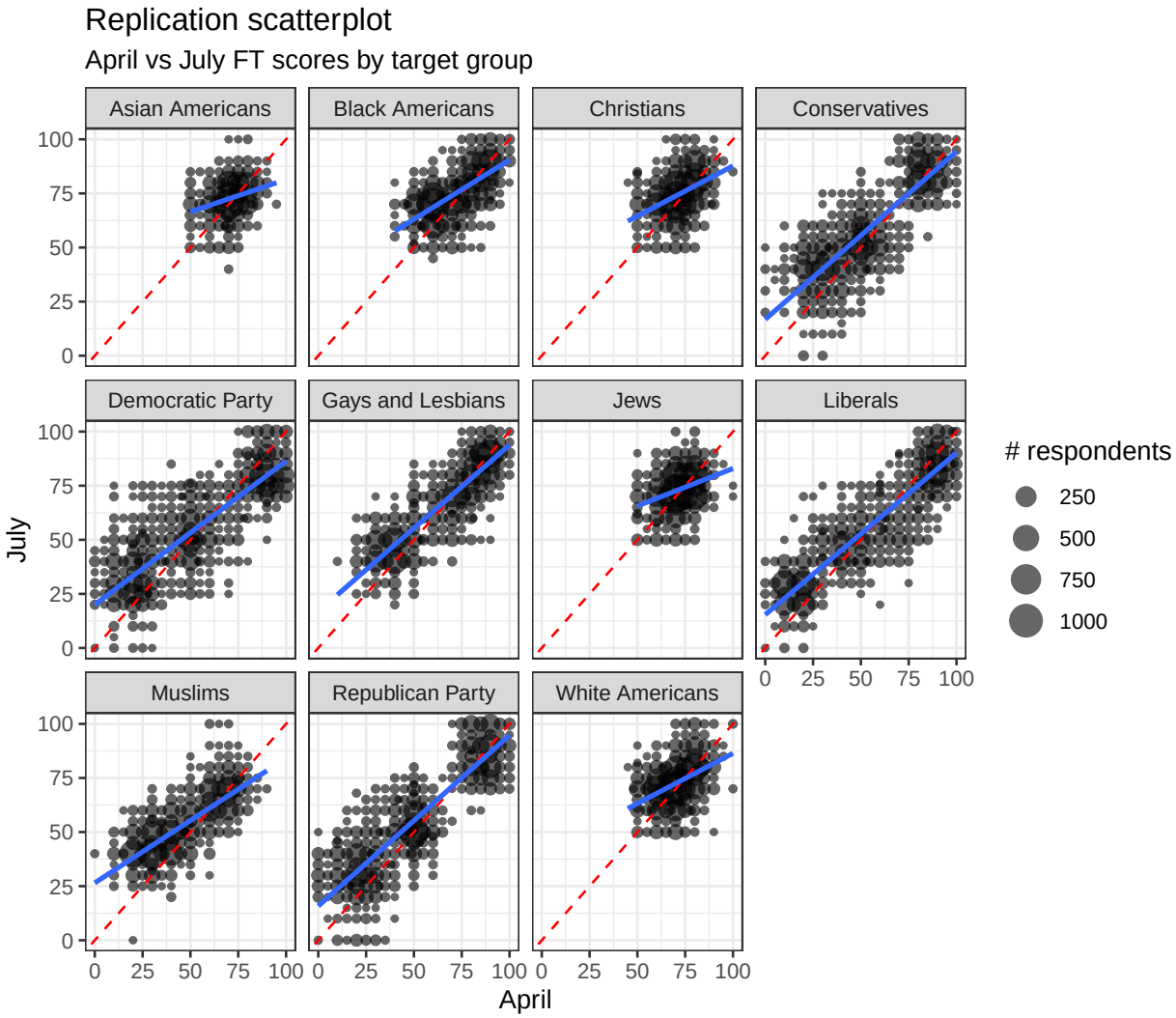

Figure 7: Feeling thermometer scores for the April (x-axes) and July (y-axes) vintages of the simple prompt, broken out by target group (facets).

the April and July runs of the same prompt. From this analysis, it would appear that the evidence of moderation in the aggregate data is more a function of weak or zero correlations than of an actual reduction in extremism. Put differently, Figure 8 further confirms our conclusion that the synthetic data is driven primarily by politics.

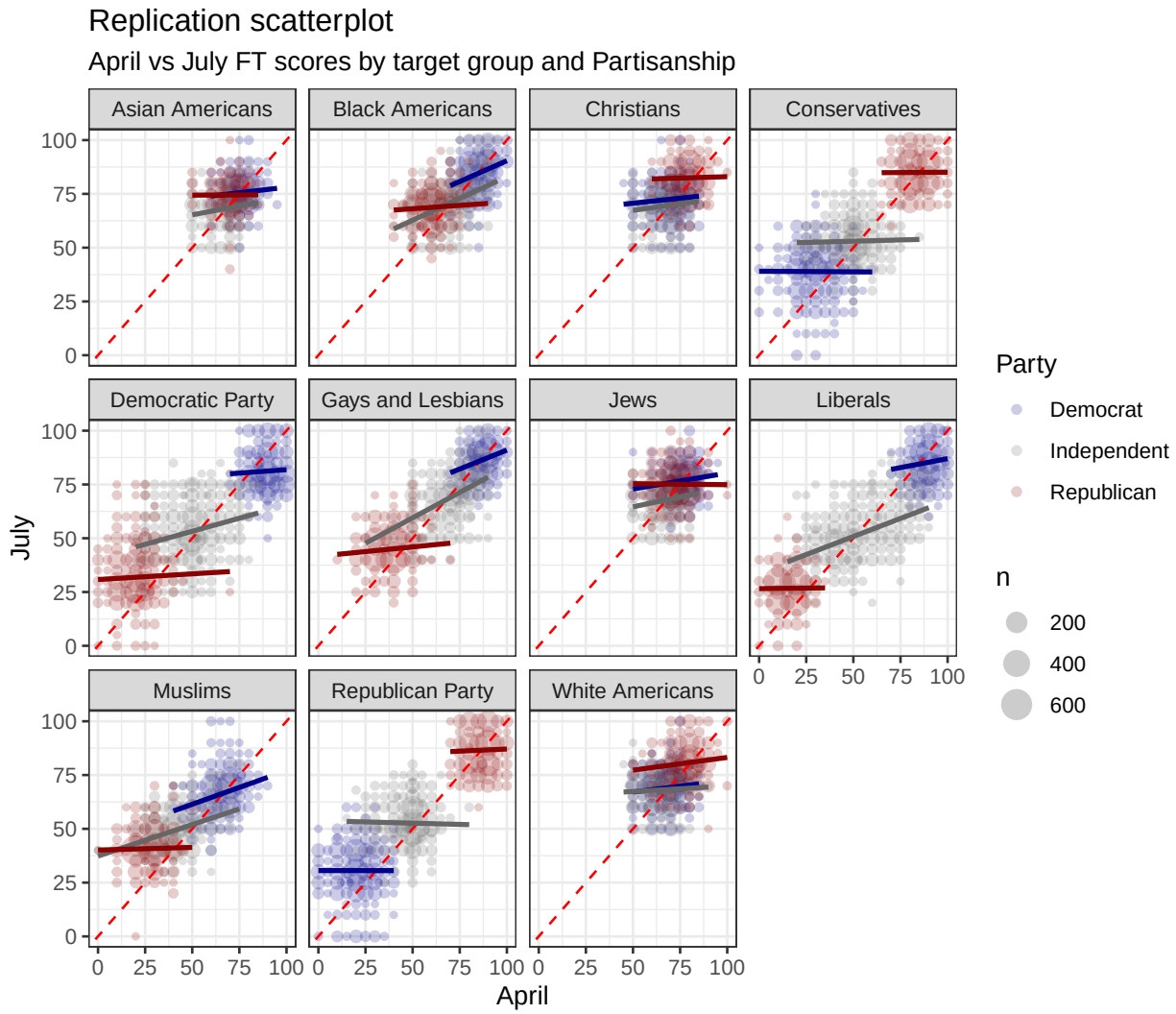

Figure 8: Feeling thermometer scores for the April (x-axes) and July (y-axes) vintages of the simple prompt, broken out by target group (facets) and party ID of synthetic respondents (blue circles are Democrats, grey circles are Independents, and red circles are Republicans).

5 Validation

Our main results relied on ChatGPT's numeric responses to a prompt about expressing feelings in terms of numbers between zero and 100. However, numbers may be a more difficult unit for a language model to produce, since they exist in a similar embedding space, meaning that the LLM is more likely to draw from them at random. This is due to the fact that tokens in close proximity in the embedding space are also proximate in terms of their posterior probability. To the extent that this is a concern, the fact we do observe consistent patterns across covariates is reassuring. Nevertheless it is worthwhile to validate these values.

To do so, we prompted ChatGPT to provide a short explanation for why it chose the feeling thermometer number it did. There was surprising consistency in these explanations, to the extent that we can document the correlation between a truncated version of the explanation (removing references to a particular group and non-alphanumeric characters) and the feeling thermometer score. Figure 9 plots the most commonly occurring explanations (those that are used 100 or more times) and their associated feeling thermometers, sizing points by the number of instances of each explanation-FT score pair.

The plot provides very reassuring validation of the numeric feeling thermometer scores in two ways. First, a cursory glance at the y-axis highlights that explanations associated with warmer and colder FT scores are sensible ("i have a very warm feeling", "i feel very favorable" etc. for higher scores; "i do not feel favorably", "i have very negative feelings", "i strongly disagree with their policies and beliefs" etc. for lower scores). Second, there is much more consistency vertically than horizontally, meaning that a given explanation is only associated with a handful of clustered FT scores, whereas a single FT score might have several different types of explanations. The exceptions to this latter pattern are also intuitive. We document a wider range of FT scores associated with "i don't feel particularly warm or cold" and "i have mixed feelings" than the scores associated with "i have very negative feelings" and "i feel very warmly".

To more robustly validate the feeling thermometer scores produced by ChatGPT, we pursue three different extensions. First, we use LDA [Blei et al., 2003] to generate topics for each explanation. We estimate 30 topics (k) using both unigrams and bi-grams on the explanations, producing a matrix of document-topic weights $\theta_{d,k}$ describing how likely it is that explanation d is about topic k, as well as a matrix of word-topic weights $\phi_{w,k}$ that describe how likely it is that word w is associated with topic k. We use these values to calculate the weighted average feeling thermometer score per topic by weighting each synthetic response by the topic distribution θ across documents (explanations). Formally:

$$\text{Avg FT score}_k = \sum_d \theta_{d,k} * \text{FT score}_d \tag{1}$$

As illustrated in Figure 10, there is a reasonable association between the topics (indicated on the y-axis by the top 5 most associated words, determined by the ϕ word-topic distribution) and the feeling thermometer scores most commonly found in the explanations that are more heavily associated with the topics. For example, the warmest feeling thermometer scores are those associated with the topic concerning equal rights and support for members of the LGBTQ+ community when used by Democrats. Conversely, the coldest topic is described by terms like "disagree" and "cold", as well as the bigram "disagree policies". Interestingly, the topic associated with conservatives ("conservatives", "towards_conservatives", "values", "conservative", "government") is associated with cold FT scores among Democrats, but warm FT scores among Republicans.

Second, we use a BERT transformer to embed these explanations in semantic space, producing numeric representations of each explanation in a 768-dimension vector. If the AI is using feeling thermometer numbers accurately, we would expect that the difference between the explanations

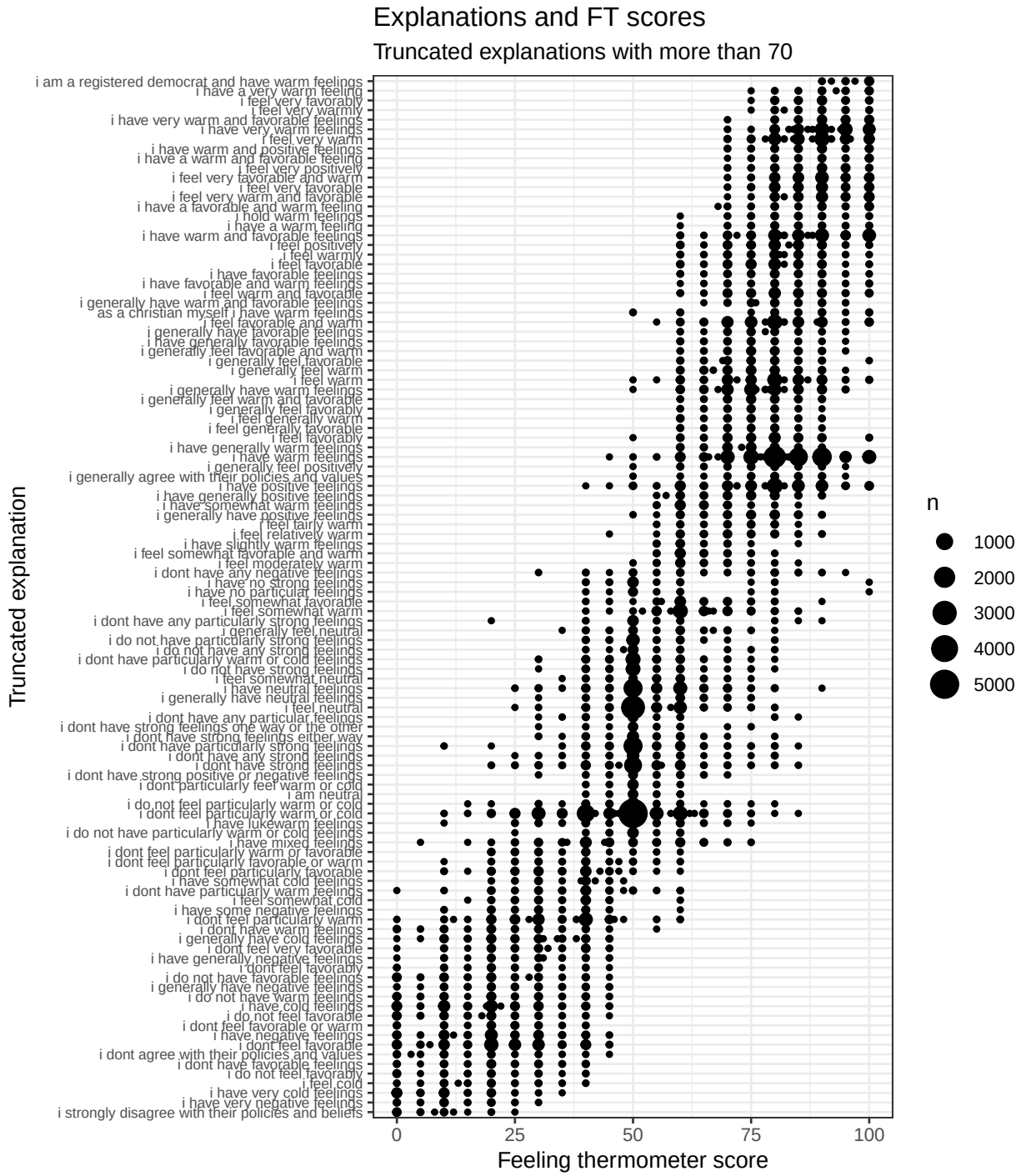

Figure 9: Feeling thermometer scores (x-axis) by truncated explanation (y-axis).

associated with a 10 score and a 20 score are similar to those between an 80 and a 90 score, and that these are both smaller than the differences between a 10 score and an 80 score. We can measure the difference between feeling thermometer scores with a simple absolute difference metric. To capture the difference between the explanations, we rely on a cosine similarity metric of each explanation's 768-element long vector representation. To simplify the analysis, we aggregate

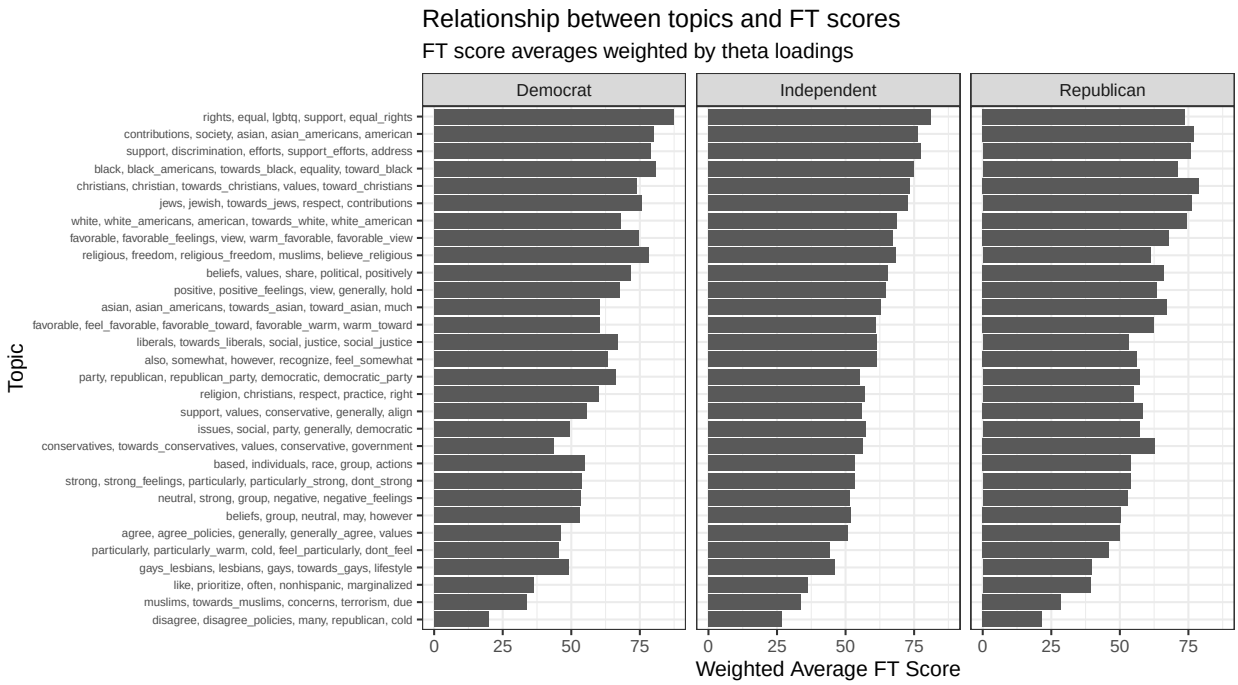

Figure 10: Weighted average feeling thermometer score (x-axes) for thirty different topics (y-axis) estimated on explanations provided by ChatGPT 3.5 turbo for why it chose the feeling thermometer it did, disaggregated among Democrat, Independent, and Republican prompts (columns). The weighted average uses the probability that topic k is associated with explanation d to weight the contribution of explanation d's associated feeling thermometer score. Formally: Weighted Avg FT score$_k = \sum_d \theta_{d,k} * \text{FT score}_d$.

our data twice. First, we calculate the average vector representation for each feeling thermometer score-by-target group-by-sociodemographic profile (dropping rarely occurring FT scores which turn out to be those not rounded to 5 unit increments). Then, after calculating the cosine distance for all vector representations, we then collapse the data once again to the average cosine distance for each feeling thermometer distance-by-party ID. Figure 11 visualizes the results with the difference in feeling thermometer responses on the x-axes and cosine distances between the vector representations on the y-axes, with each facet indicating a target group-by-party ID.

As illustrated, the plots trend upward in almost every facet, albeit never linearly, indicating that numeric feeling thermometer responses which are further apart are accompanied by explanations which are similarly further apart in semantic space. For reference, an example of two explanations that are far apart in semantic space are "I feel warm toward Christians. I strongly identify as a Christian and believe in their values and teachings." (FT score = 90) versus "I feel very cold toward the Democratic Party. They do not represent my values, and I disagree with their policies." (FT score = 0). This Euclidean distance between their thermometer scores of 90 corresponds to a cosine distance between their vector representations of roughly 0.81 on a support ranging between 0 (semantically the same) to 1 (semantically totally dissimilar).

Finally, we use ChatGPT itself to assist us in annotating a random sample of 2,000 pairs of explanations.[3] In each comparison, we ask ChatGPT 3.5 to indicate which explanation justifies a warmer evaluation of some group. We then compare these labels to the true difference in FT scores,

[3]This idea was inspired by Wu et al. [2023].

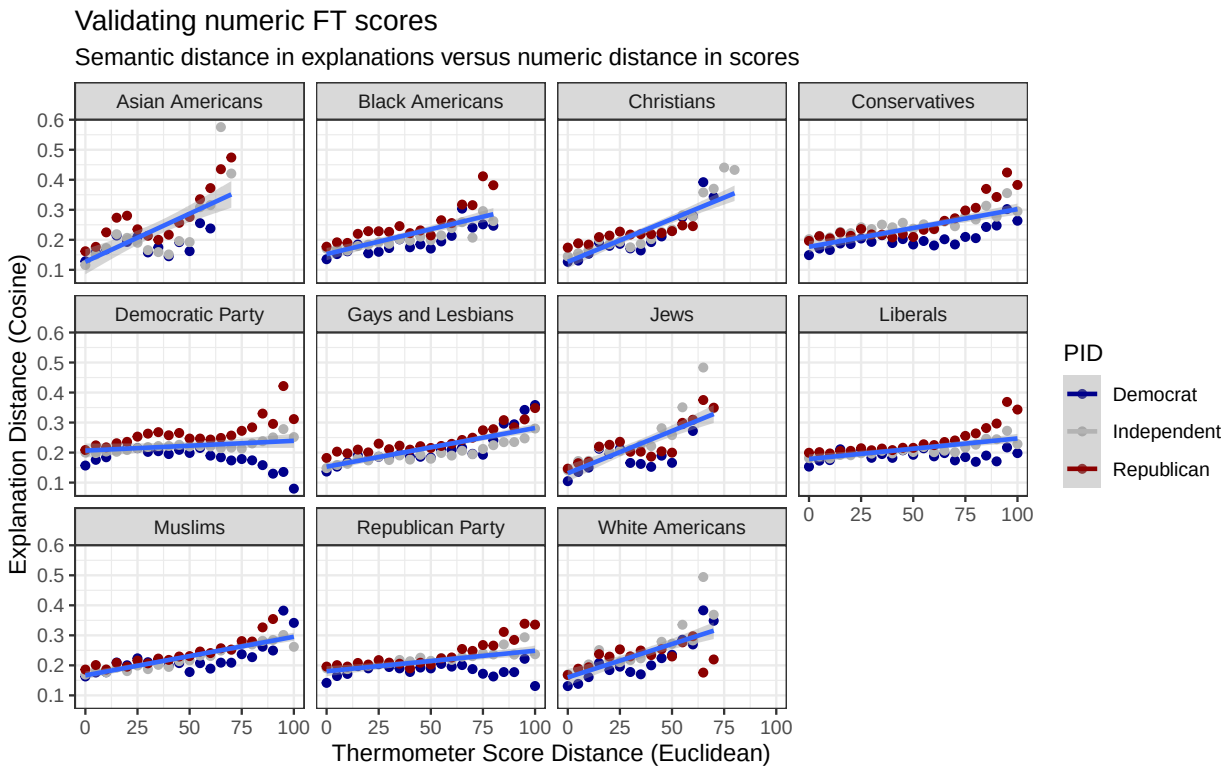

Figure 11: Comparison between how far apart two feeling thermometer scores are (x-axes) compared to how far apart their explanations are in 768-dimensional embedding space using cosine distance (y-axes). Points are colored by the party ID of the prompt and disaggregated by the target group (facets).

finding very strong performance overall as indicated by the confusion matrix in Table 2, with an F-1 score of 0.81. Furthermore, Figure 12 suggests that disagreements between the AI's choice of more positive explanation and the associated FT score difference occur closer to zero – i.e., where the feeling thermometer scores are more similar.

We further confirm this result by applying the Bradley-Terry model described in Wu et al. [2023] to calculate the latent dimension of warmth. In brief, this method assumes that the probability that an actor i "wins" in a pairwise comparison against actor j is simply the ratio of i's latent "ability" α_i to that of j. Formally:

$$Pr(i > j) = \frac{\alpha_i}{\alpha_j} \tag{2}$$

By setting $\lambda_i = exp(\alpha_i)$, estimation if these latent abilities is easily done via a logistic regression via

$$log\left[\frac{Pr(i > j)}{Pr(j > i)}\right] = \lambda_i - \lambda_j \tag{3}$$

Following Wu et al. [2023], we estimate these "ability" scores (or in our setting "warmth" scores) using the '`BradleyTerry2` package for `R`, and set a score of 50 as the reference point. As illustrated in Figure 13, the latent measure of warmth generated by the Bradley-Terry model recovers the feeling thermometer scores reasonably well, with scores below 50 being associated with lower warmth scores, and those above 50 being associated with higher warmth scores.

	#1 warmer	#2 warmer
#1 warmer	732	217
#2 warmer	190	861

Table 2: Confusion matrix summarizing results of pairwise comparison between two explanations, one of which was associated with a warmer feeling thermometer score than the other. Rows indicate predictions of which of two explanations is warmer according to ChatGPT 3.5-turbo when asked to read both explanations, and columns indicate the truth. F-1 score = 0.81.

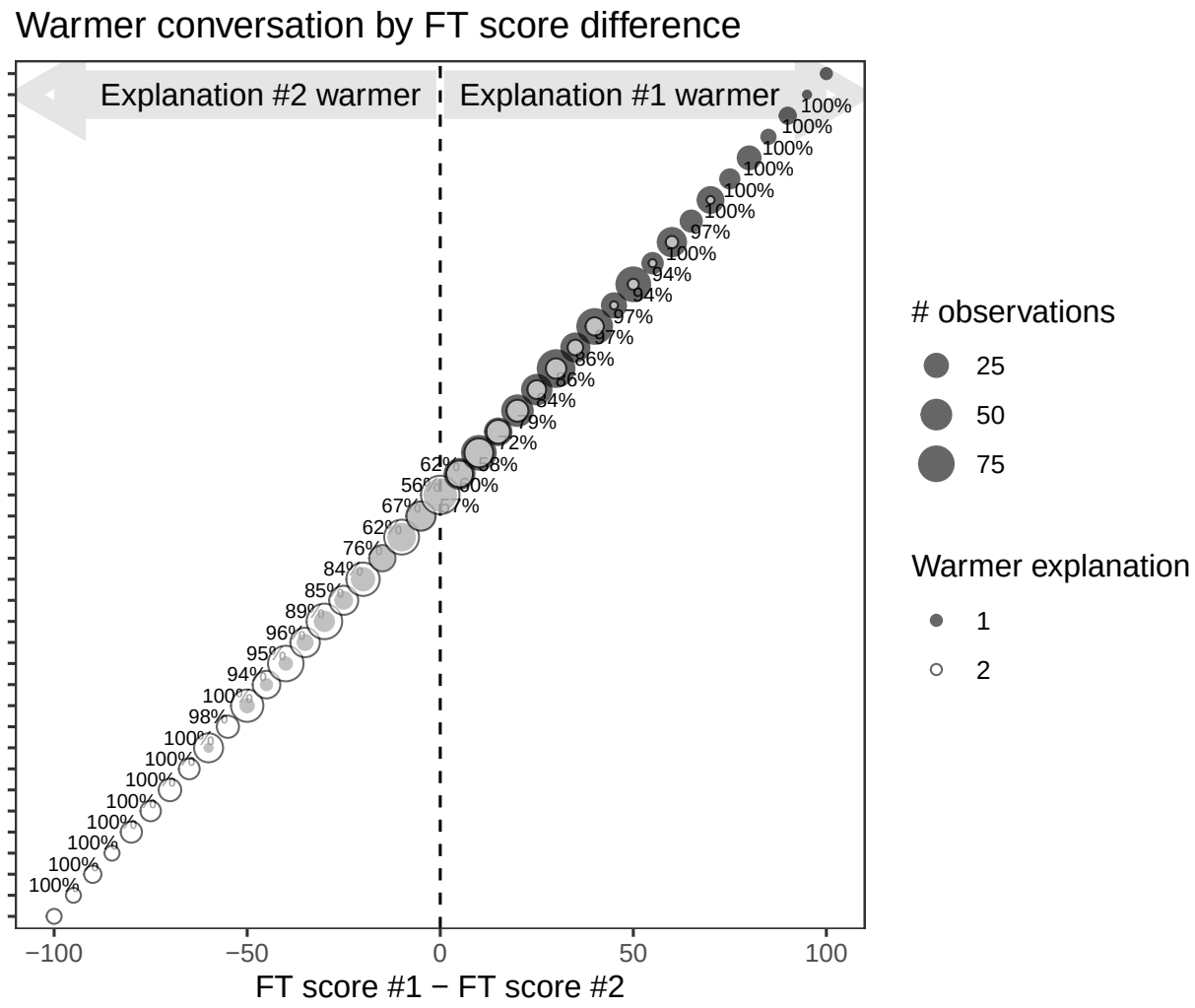

Figure 12: Each point represents a difference between two feeling thermometer scores chosen at random (x-axis), rounded to the nearest 5 and arranged in order from comparisons in which the first score is warmer than the second (y-axis indicates order and is included only to aid visual clarity). Each of the 2,000 total pairwise comparisons was evaluated by ChatGPT 3.5-turbo to determine which explanation expressed warmer sentiments, the results of which are indicated by either solid (if the AI concluded that the first explanation was warmer than the second) or hollow (if the AI concluded that the second explanation was warmer than the first) points. Points are sized by how many of the 2,000 randomly selected scores and comparisons fell into the bin rounded to the nearest 5 point gap. Labels indicate the proportion of comparisons that were coded correctly by the AI. When the feeling thermometer scores are closer together (i.e., the x-axis is closer to zero), the explanations are harder to distinguish.

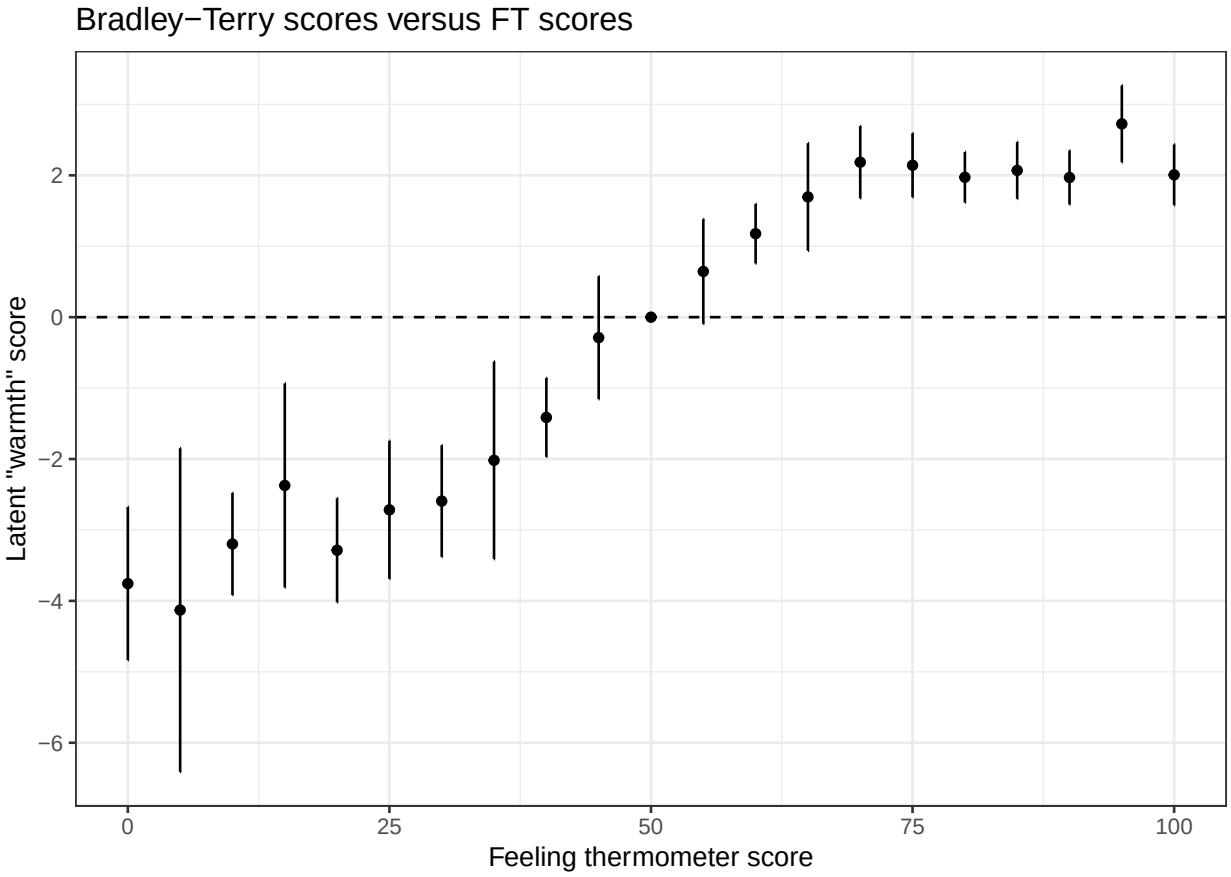

Figure 13: Feeling thermometer scores (binned to 5-unit intervals, x-axis) against "warmth" scores (y-axis) estimated using the Bradley-Terry method of measuring a latent trait in pairwise comparisons.

6 Initial Prompt

In the process of writing this paper, we have adjusted our prompts, finding differences in the synthetic data generated by the same prompt between when we initially collected our data in April 2023, and when we re-ran the API in July of 2023. These changes have altered a core conclusion drawn in our initial analysis: namely that synthetic data exaggerated the evidence of affective polarization in the United States. These results do replicate in several versions of our subsequent synthetic data collections (notably not in the first-person prompt, however), but are inconsistent across persona covariates and target groups. As such, we have removed these results from the focus of our manuscript, and reproduce them here for the sake for posterity.

6.1 Exaggerated Extremism

The main conclusion from our initial analysis was that the synthetic data exaggerated the evidence of out-group antipathy along the dimensions of partisanship and ideology by a factor of approximately 6.5. Substantively this means that the gap between feeling thermometer scores toward Democrats and Republicans was approximately 6.5 times larger in the synthetic data based on the April 2023 run of the simple prompt, compared to the same measure calculated in the ANES data. We visualize this exaggerated extremism in Figures 14, 15, and 16 below.

Figure 14 indicates that synthetic Democrats and Republicans feel warmer toward their co-ideologues, and colder toward their out-ideologues, than real humans do. Similar exaggerations are documented in attitudes toward the political parties, although these only obtain for the in-group parties. Out-group evaluations measured in the synthetic data are similar to those measured among human respondents.

Figure 15 documents similar patterns by calculating the difference between two societal groups (conservatives and liberals, Republican and Democrat parties, white and black Americans, and Christians and Muslims) in both the synthetic and human samples. We then plot these differences by race (rows) and party affiliation (y-axes), where zero indicates that there is no difference in feeling thermometer scores between the two groups. Bars that lie further from zero indicate greater polarization estimated in that data. As illustrated, in almost every comparison, the polarization measured in the synthetic data is larger than that found in the human sample.

A final test confirms that not only are these differences larger in the synthetic data than in the human data, but that these differences are statistically significant in all but a small minority of examples. Figure 16 visualizes these results as the difference in the thermometer gap described above, between the synthetic data and the ANES. Positive values (in black) indicate how much larger the gap is in the synthetic data, while negative values (in red) indicate how much larger the gap is in the human data. As illustrated, across all race-by-partisan-by-gap profiles, the ANES data is more polarized in only 5 examples, and is only significantly different from the synthetic estimate in two of these. The rest all point toward the synthetic data exaggerating the thermometer gaps, and significantly so in all but two cases (Hispanic independent measures of the partisan gap, and Black Republican measures of the racial gap).

6.2 Replicating with different survey

The preceding conclusions rely on ANES data, one of the most well-known nationally representative public opinion polls of U.S. politics. We also validated our initial results using a bespoke survey of 2,322 online respondents fielded in 2021 that investigated affective polarization in the context of the Covid-19 pandemic (Clinton and Kam [2022]). While the number of target groups is reduced in

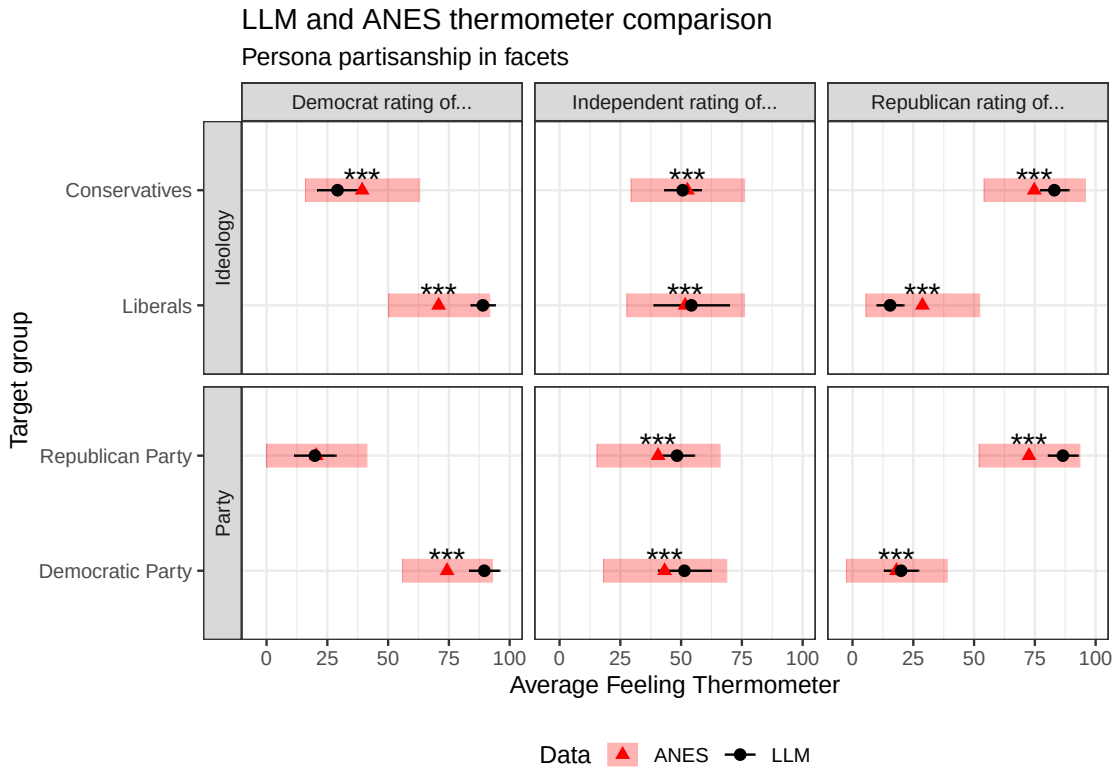

Figure 14: Average feeling thermometer results (x-axis) for different target groups (facets) by party ID of respondent (y-axis). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical. Statistically significant differences indicated with *** = $p < .001$; ** = $p < .01$; * = $p < .05$.

this survey, we nevertheless are able to implement the same methods described in our manuscript, validating the generalizability of our results to a different period using a different sample from a different polling source. Figures 17, 18, and 19 recreate the figures summarized in the preceding section using this alternative dataset, finding substantively similar evidence of exaggerated feeling thermometer gaps across the dimensions of the political parties and whites versus black Americans.

6.3 Detailed description of summary statistics

To calculate the degree to which ChatGPT-generated responses are more extreme than found among human respondents in the ANES, we turn to our matched data where every human respondent in the ANES is matched with a set of random pulls from the ChatGPT API, prompted to adopt the persona of the ANES respondent along the dimensions of age, gender, race, education, income, and partisanship. For each demographic profile defined by these characteristics, we calculate the average feeling thermometer expressions toward the set of target groups asked in the ANES. We then calculate the same average among the synthetic humans generated by ChatGPT which were matched to the real ANES respondents. For each demographic profile, we thus obtain an average feeling thermometer for a given target group estimated by ChatGPT and among the humans

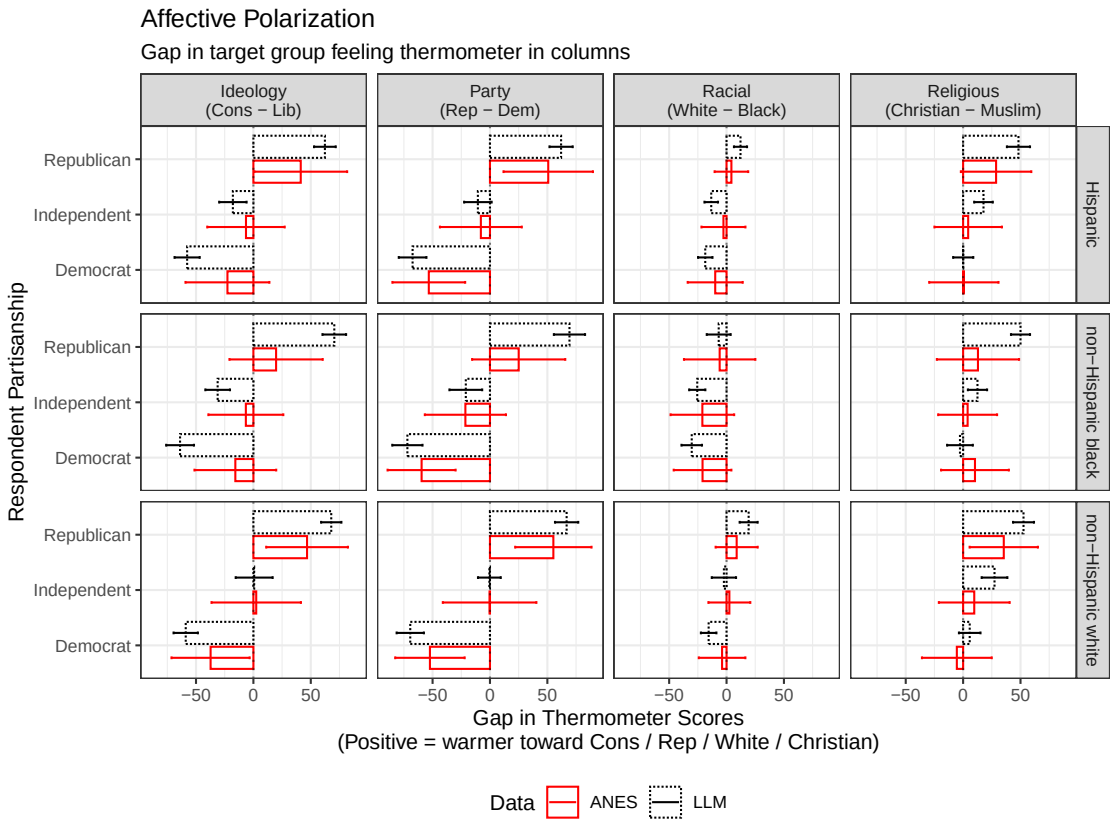

Figure 15: X-axes measure the difference in feeling thermometer ratings between two target groups (columns), by the party ID of the respondent (y-axes) and their race (rows). Black dotted lines indicate results generated by synthetic respondents from the LLM. Red solid lines indicate results generated by real humans from the ANES. Horizontal lines indicate one standard deviation.

sampled by the ANES. Since a thermometer score of 50 captures an indifferent or "neutral" attitude toward a given outgroup, we measure extremism as the absolute difference between the recorded attitude and 50. Our summary statistic of ChatGPT's extremism is thus the ratio of the LLM model's average absolute difference divided by the ANES data's average absolute difference. Ratios greater than 1 indicate that ChatGPT's estimates are more extreme (either more warm or more cool) than real human attitudes, while those less than 1 indicate the opposite. On average, across all covariate profiles and all target groups, synthetic responses gathered from ChatGPT are 4.88 times more extreme than those recorded among real human respondents to the ANES. Broken out by target group, we see consistent evidence that ChatGPT's estimates are always further from 50 on average than the averages found among real humans responding to the ANES.

Turning to the question of antipathy toward outgroups (measured as either affective polarization, partisan sectarianism, or racial or religious antipathy), we pursue a similar exercise. Specifically, we calculate the measure of "polarization" along each dimension, subtracting sentiments towards liberals, Democrats, Blacks, or Muslims from the sentiments towards conservatives, Republicans, Whites, and Christians, respectively. The resulting measures are positive when the respondent is more warm toward stereotypically conservative / Republican groups, and negative when the respondent is more warm toward stereotypically liberal / Democratic groups. Given the matched nature of the data, we are able to calculate these measures using both the real human's

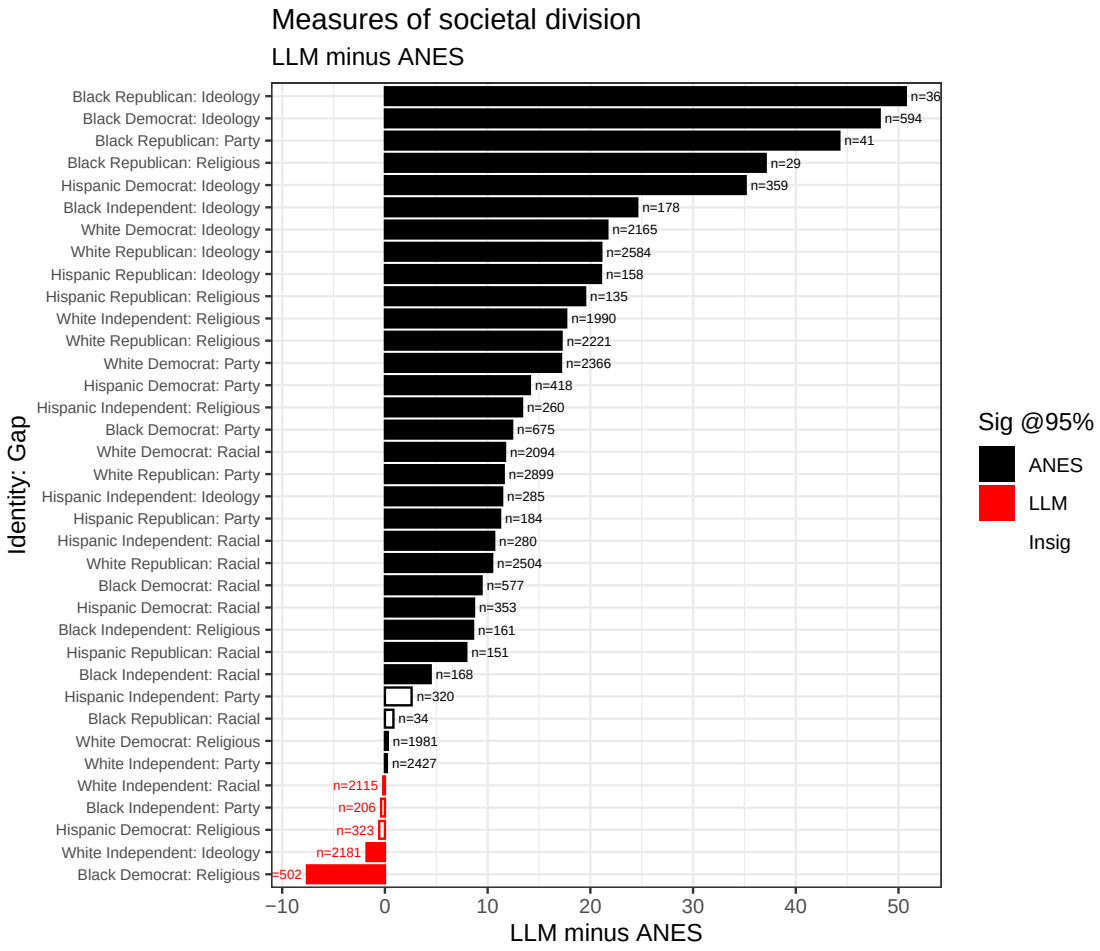

Figure 16: Difference in estimated societal polarization along the dimensions of ideology, partisanship, race, and religion between thermometer gaps estimated from ChatGPT (LLM in black) and from human survey respondents (ANES in red), by race and partisanship of the respondents (y-axis). Solid bars indicate differences between the two data sources that are significant at the 95% level of confidence, while hollow bars indicated statistically insignificant results at this threshold. Number of respondents in each category with ANES responses given by numbers.

answers to ANES survey questions as well as what ChatGPT thinks someone fitting their profile would say. We then calculate the ratio of the absolute value of the LLM's measure of polarization relative to that found in the ANES for each respondent (again relying on the absolute value). The resulting overall average ratio across all respondents and all groups is just under 7, suggesting that ChatGPT estimates are biased toward out-group antipathy.

However, ratios are sensitive to outliers, particularly in the denominator. If the ANES difference happens to be zero (a likely event given the pressures of social desireability bias in which respondents want to be seen as egalitarian), whatever small difference found in the LLM results will be exaggerated. We therefore plot the raw distributions of polarization across the four measures of partisanship, ideology, race, and religion (y-axis) in Figure 21. As illustrated, ChatGPT's estimates are consistently more polarized than those found among human respondents to the ANES, although we note that there are many examples where an ANES respondent's attitudes are more polarized than its synthetic counterparts.

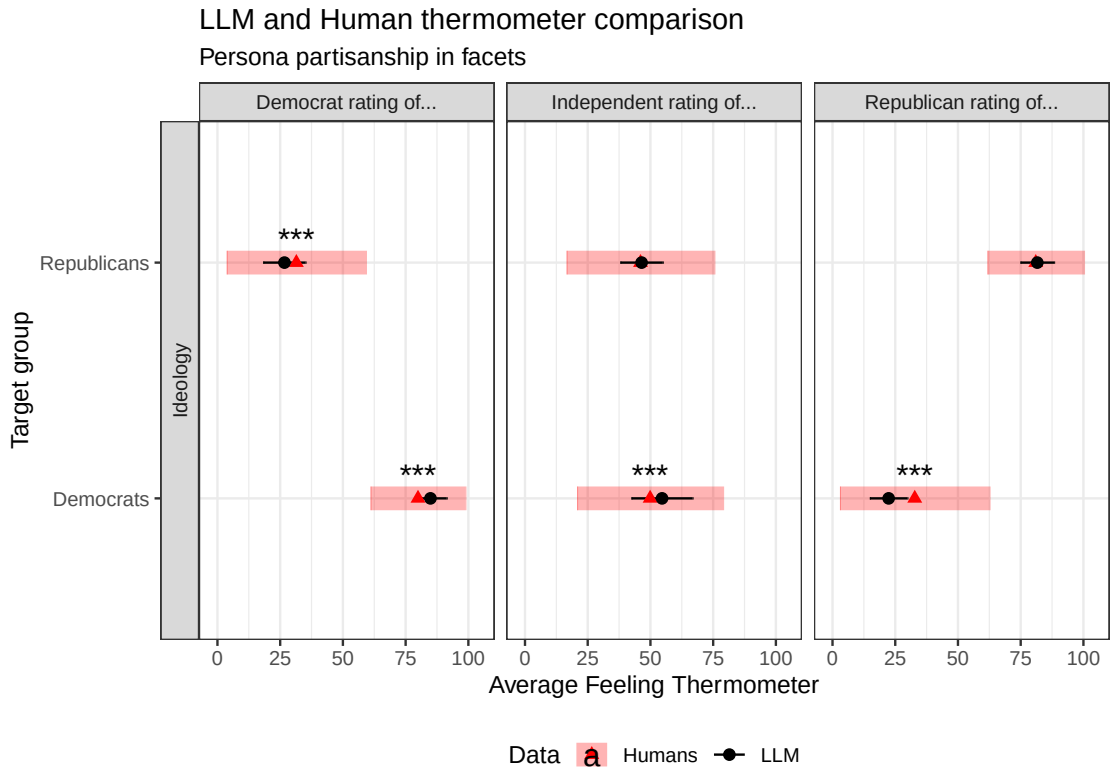

Figure 17: Average feeling thermometer results (x-axis) for different target groups (facets) by party ID of respondent (y-axis). Average human estimates from bespoke 2021 survey on affective polarization during Covid-19 indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical. Statistically significant differences indicated with *** $= p < .001$; ** $= p < .01$; * $= p < .05$.

To evaluate the degree to which ChatGPT estimates are less variable (i.e., more confident) than estimates generated by real humans, we start by calculating the standard deviation for each race-by-party-by-target group profile in the LLM and ANES data, then divide the latter by the former. We plot the results of this exercise in Figure 22, averaging over race and party ID to calculate a summary measure of how much more confident ChatGPT is in its estimates than the ANES. As above, the vertical dashed line at 1 indicates parity. As illustrated, ChatGPT's standard deviations are consistently between one-quarter and one-third the size of those found among human respondents to the ANES, with an overall average of roughly 0.314.

An alternative way of characterizing ChatGPT's exaggerated precision is by looking across all demographic profiles with two or more respondents in the ANES data, and dividing the ChatGPT standard deviation by that in the ANES. We plot each profile's ratio in Figure 23, sizing the points by the number of observations in each profile, and labeling each outcome by the proportion of profiles whose ChatGPT-derived estimate is more precise than the ANES-derived estimate. Overall, 95% of profiles with two or more ANES respondents that expressed a feeling thermometer toward an outgroup had smaller standard deviations when estimated using ChatGPT relative to human respondents to the ANES. And as Figure 23 makes clear, this conclusion is far stronger if we focus on demographic profiles with more ANES respondents.

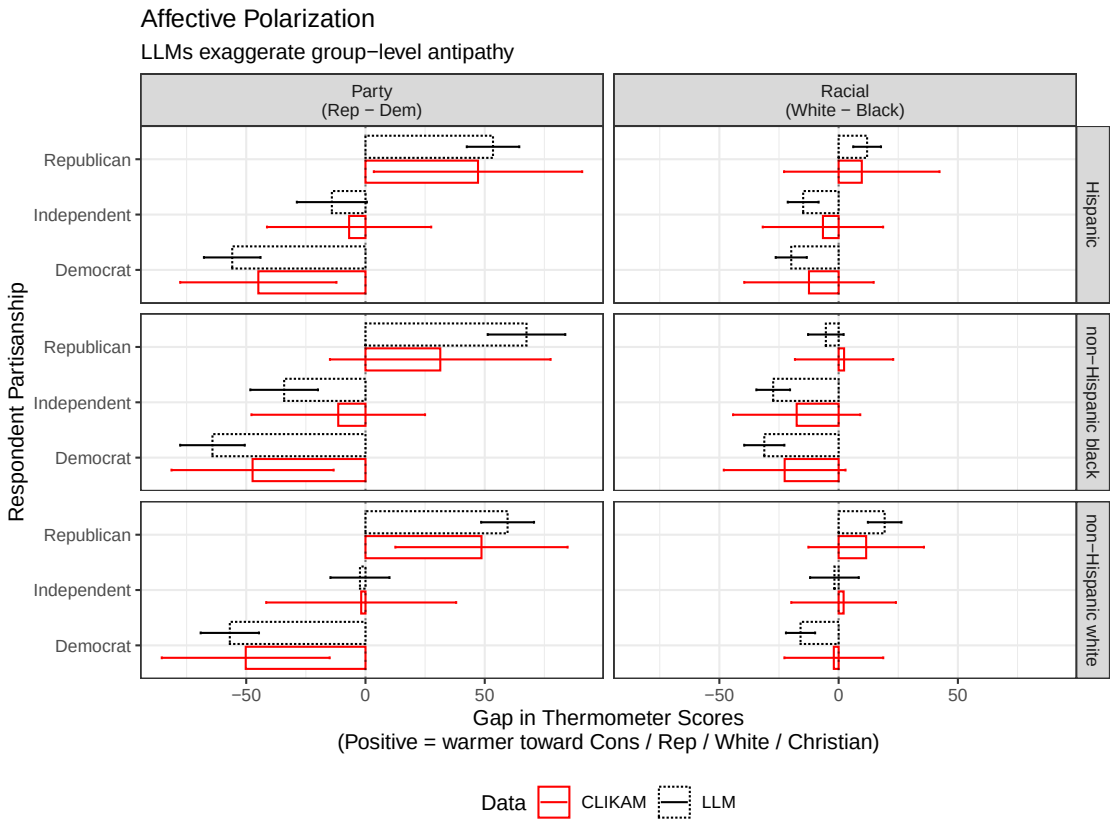

Figure 18: X-axes measure the difference in feeling thermometer ratings between two target groups (columns), by the party ID of the respondent (y-axes) and their race (rows). Black dotted lines indicate results generated by synthetic respondents from the LLM. Red solid lines indicate results generated by real humans from bespoke survey on affective polarization during Covid-19. Horizontal lines indicate one standard deviation.

6.4 Concluding the evidence of exaggerated extremism

The preceding findings were all based on the April 2023 version of the simple prompt, described above in Table 1. Subsequent runs of the same prompt, as well as the first-person modification, and the more detailed version of the prompt, yield less convincing evidence of this particular dimension along which the synthetic data fails to recover real human patterns. Empirically, this attenuation in our findings is likely due to the evidence of mean reversion over time, as documented in our manuscript. However, as we document in Figure 24, the core conclusion that synthetic data exaggerates our differences persists across all new samples with the exception of that generated by the first-person pronouns prompt. Consistent with the work by Levendusky and Malhotra [2016], this result suggests that asking an LLM to pretend to be someone else produces exaggerated measures of out-group antipathy, just like with humans. Nevertheless, we demonstrate in SI Section 3 that this is not a silver bullet against the myriad other issues we document with the synthetic data.

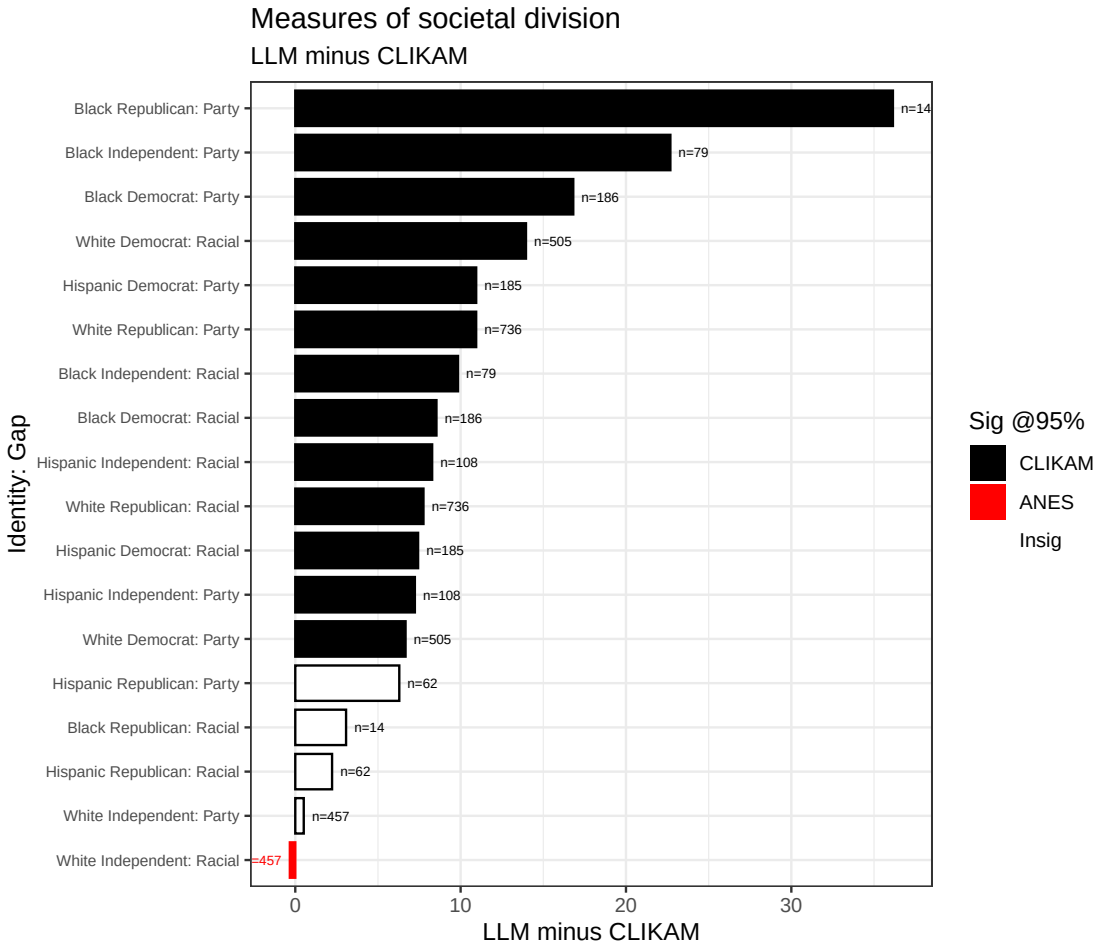

Figure 19: Difference in estimated societal polarization along the dimensions of ideology, partisanship, race, and religion between thermometer gaps estimated from ChatGPT (LLM in black) and from human survey respondents (bespoke survey in red), by race and partisanship of the respondents (y-axis). Solid bars indicate differences between the two data sources that are significant at the 95% level of confidence, while hollow bars indicated statistically insignificant results at this threshold. Number of respondents in each category with human responses given by numbers.

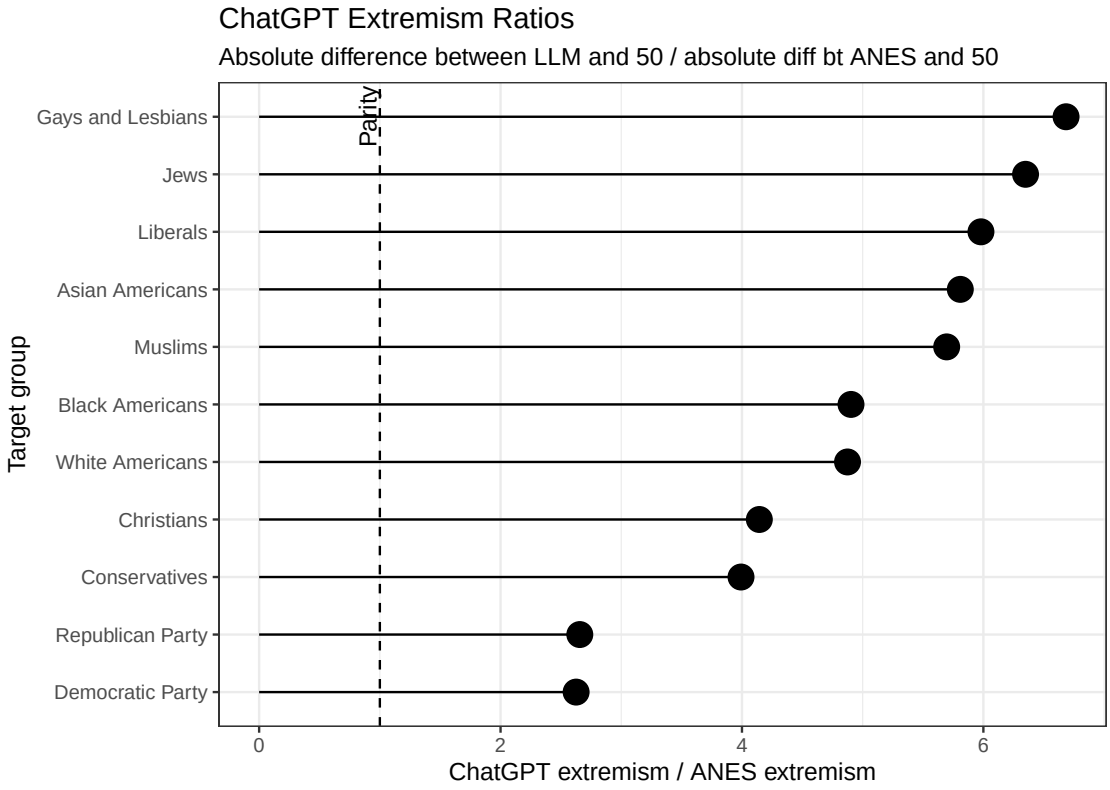

Figure 20: Ratio of ChatGPT extremism (absolute difference between estimate and 50) to ANES extremism. Dashed vertical line indicates parity.

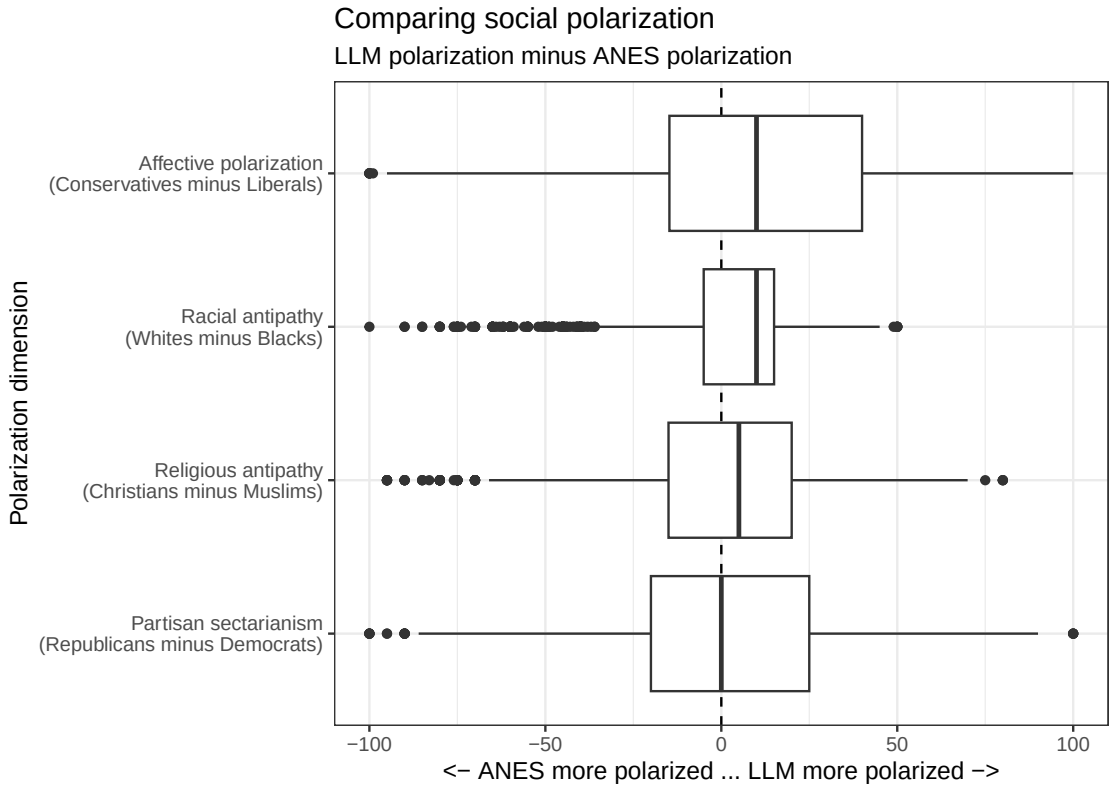

Figure 21: Difference between ChatGPT polarization and ANES polarization.

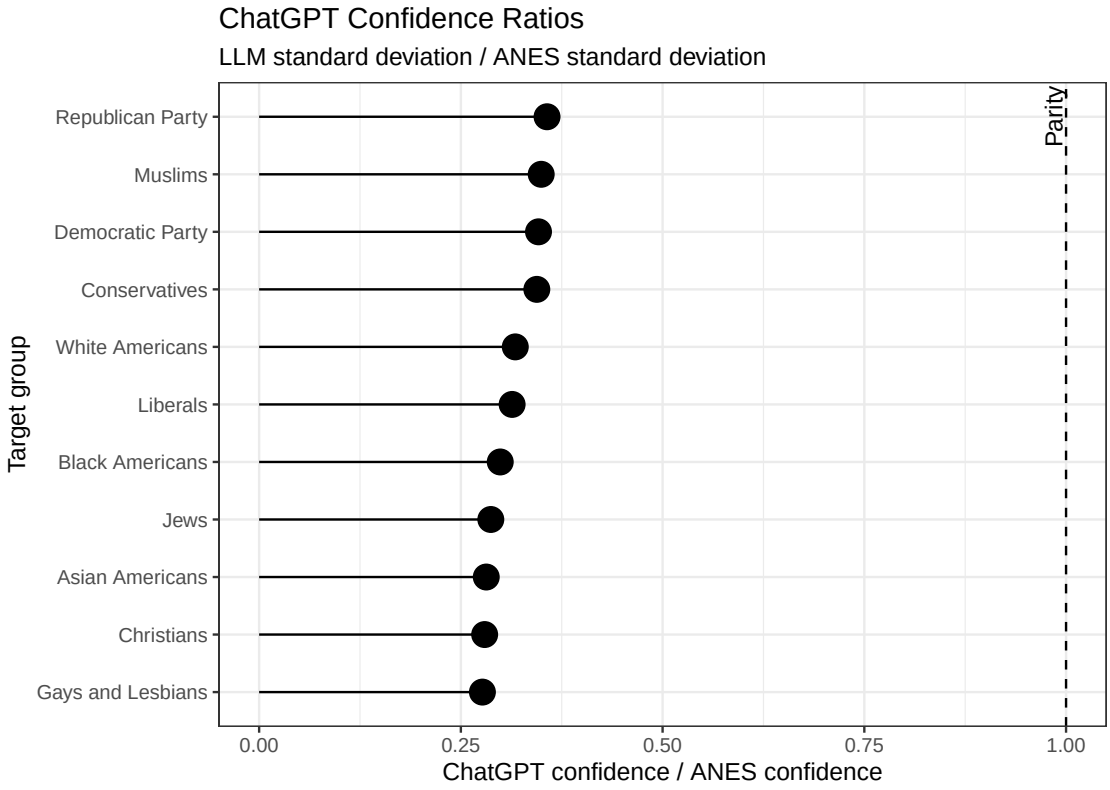

Figure 22: Ratio of ChatGPT confidence (standard deviation across race and party) to ANES confidence. Dashed vertical line indicates parity.

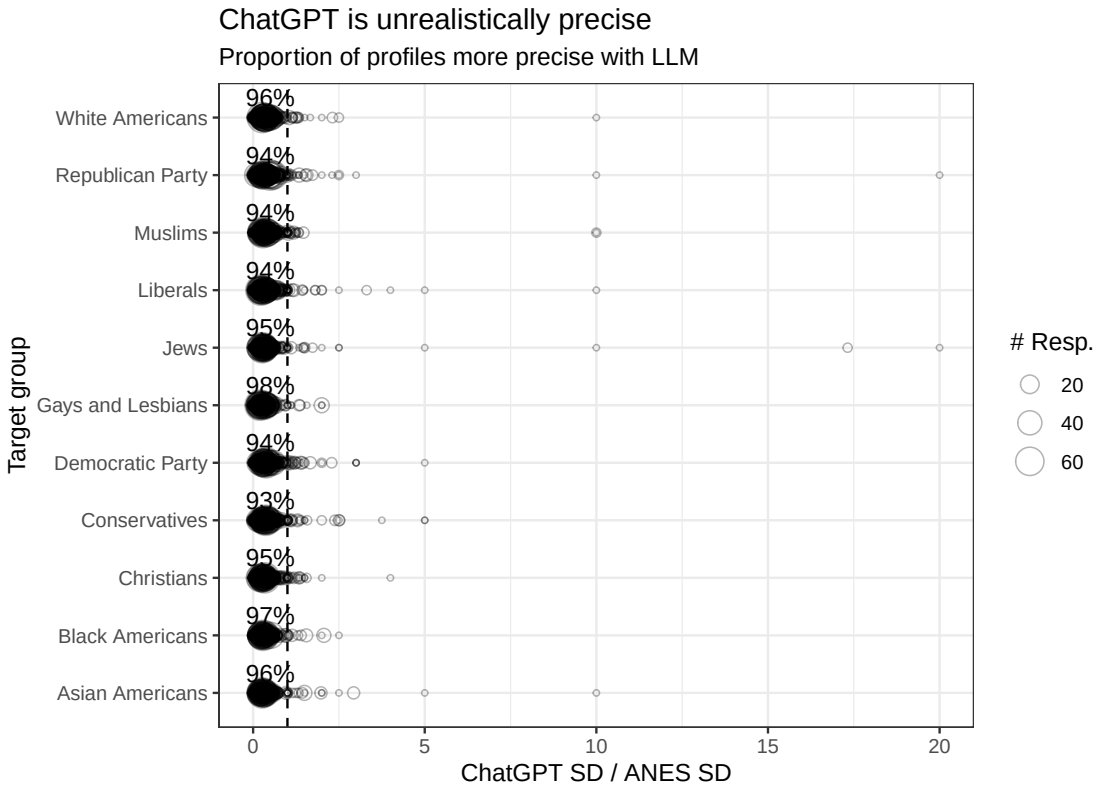

Figure 23: Ratio of LLM standard deviation to ANES standard deviation (x-axis) by demographic profile (circles) sized by total number of respondents. Ratios less than 1 indicate profiles in which the LLM-derived estimate was more precise than the same measure calculated based on the ANES. Text indicates the proportion of profiles per target group (y-axis) that were more precise when estimated via ChatGPT than via ANES.

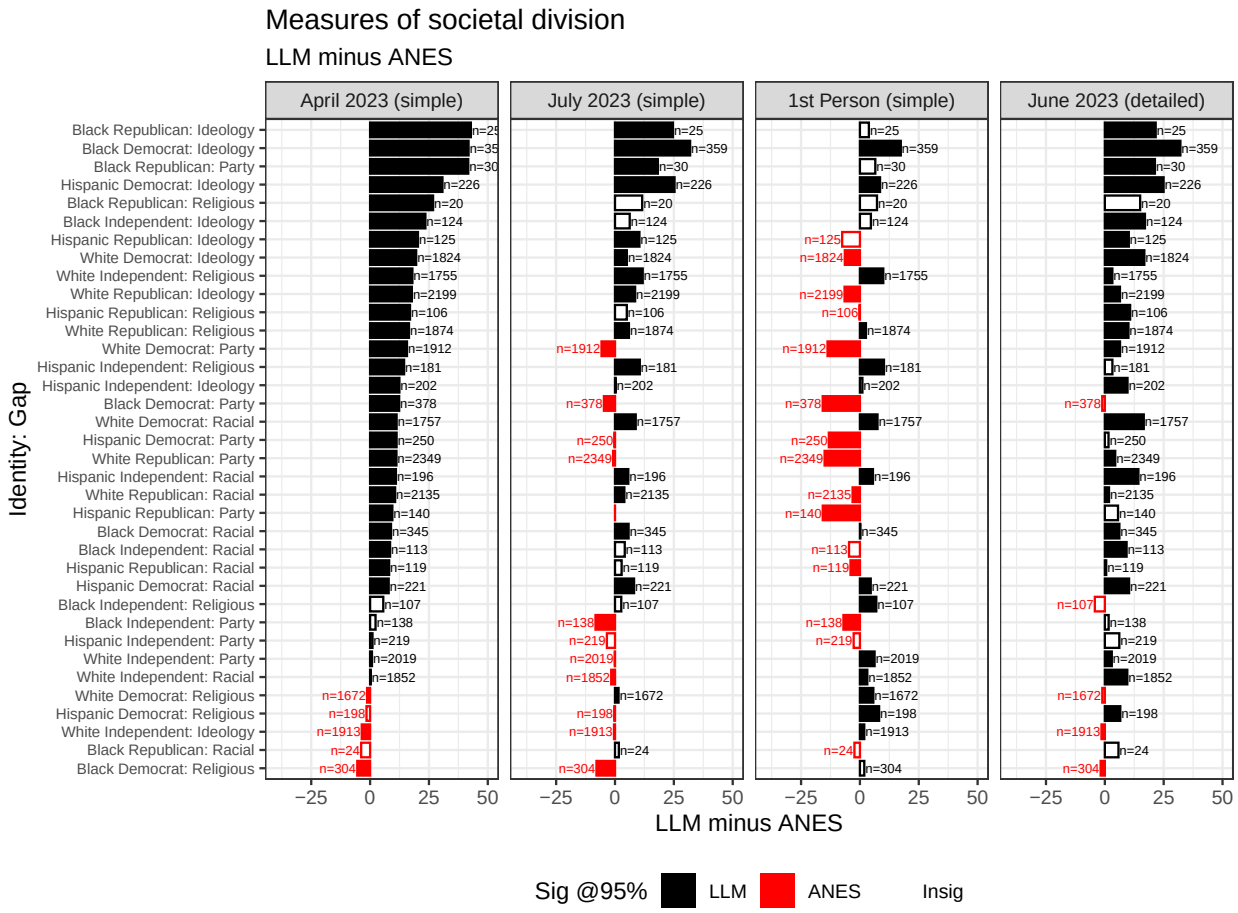

Figure 24: Difference in estimated societal polarization along the dimensions of ideology, partisanship, race, and religion between thermometer gaps estimated from ChatGPT (LLM in black) and from human survey respondents (bespoke survey in red), by race and partisanship of the respondents (y-axis). Solid bars indicate differences between the two data sources that are significant at the 95% level of confidence, while hollow bars indicated statistically insignificant results at this threshold. Number of respondents in each category with human responses given by numbers. Columns indicate different versions / vintages of the prompt.

7 "Generic" Americans

Our main results are based on prompts to ChatGPT to adopt a specific persona, defined along the characteristics of race, age, gender, education, income, marital status, ideology, interest in news and politics, voter registration status, and party ID. Our subsequent analyses then aggregated over different depths of a subset of these dimensions to characterize the bias among racial and partisan groups. Here, we instead ask ChatGPT to adopt the persona of the average American and provide the same estimates. Specifically, we instruct ChatGPT to adopt the following generalized identities:

- Basic: person, registered voter
- Party: Democrat, Republican, Independent voter
- Age: 20, 35, 50, 65 year old
- Race: non-Hispanic white, non-Hispanic black, Hispanic
- Gender: male, female
- Income: person making $30,000, $50,000, $80,000, $100,000, $150,000 per year
- Ideology: liberal, moderate, conservative

Each of these identities is not overlaid with other dimensions, meaning we only ask ChatGPT to pretend to be a person living in the USA, a 20 year old living in the USA, a Republican living in the USA, etc. We then characterize how similar these generic identities are to 1) ANES profiles and 2) the richer ChatGPT profiles used in the main analysis. In conducting this analysis, we no longer apply an exact match, and instead simply compare averages between these different sources.

Our first set of results, visualized in Figure 25, compare how similar the ChatGPT generic identities are to each other. Specifically, we calculate the average attitudes toward different outgroups with the "person living in the USA" prompt to the same generated by the "Democrat / Republican living in the USA" prompt, and then subtract the absolute value of the Democrat difference from the absolute value of the Republican difference. As illustrated, the generic American is more similar to the generic Democrat than to the generic Republican across almost all outcomes, and significantly more for groups more associated with Democrats.

To calculate the difference between generic synthetic respondents and real ANES humans, we compare the average feeling thermometer for each synthetic persona to the average feeling thermometer among actual Democrats and Republicans in the ANES. We then plot the difference in these differences in Figure 26, where negative values (meaning that the difference between the LLM and the ANES Democrats is smaller than the difference between the LLM and the ANES Republicans) are indicated in blue, and positive values are indicated in red. Substantively, the plot highlights that, across most personae and across most target groups, ChatGPT's attitudes are more similar to Democrats than to Republicans. Importantly, the only personae where this pattern doesn't hold on average are when we prompt it to adopt the identity of a conservative American, a Republican American, or a non-Hispanic white American. Perhaps even more importantly, ChatGPT's attitudes toward all target groups are more similar to actual Democrats' attitudes than Republicans, with the exception of attitudes toward Jews, the Republican Party, and Muslims. The latter, in particular, has more Republican-leaning attitudes among synthetic ChatGPT persona for every identity excepting women, Hispanics, non-Hispanic Blacks, liberals, and Democrats.

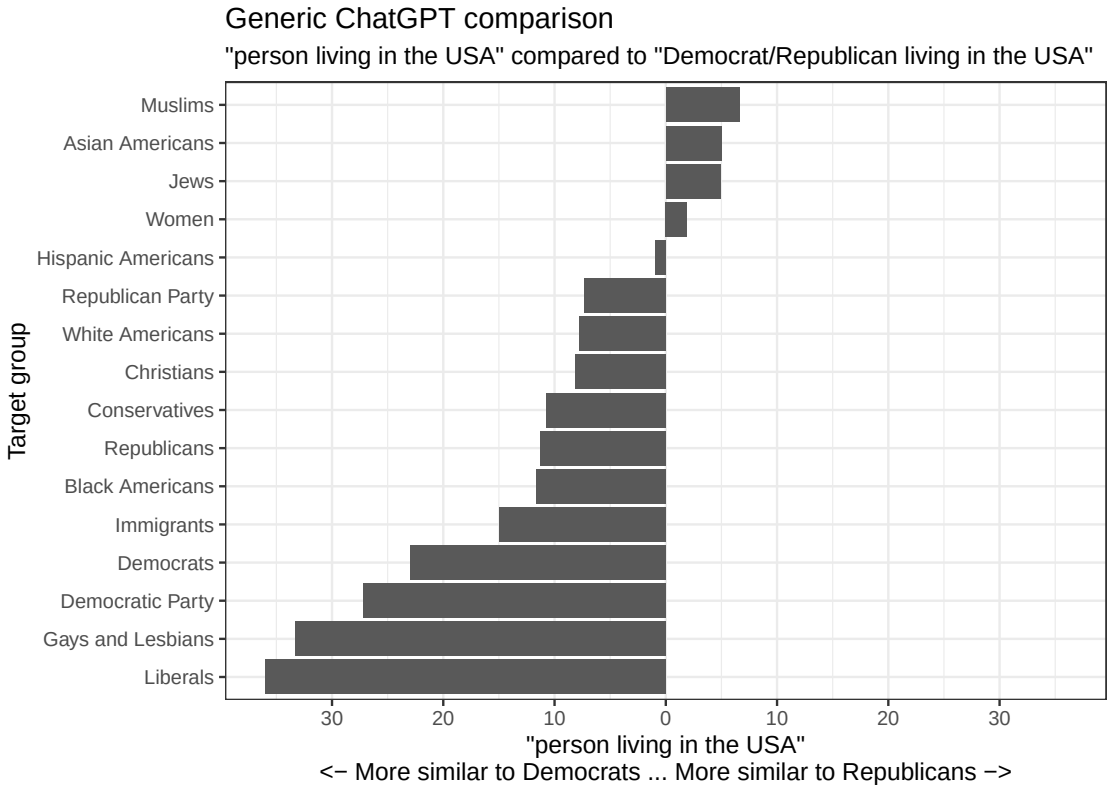

Figure 25: Comparing a generic American ("person living in the USA") with a generic Democrat / Republican, all estimated using ChatGPT. X-axis indicates the difference in the absolute gap between the generic American and the generic Democrat, and the absolute gap between the generic American and the generic Republican. Negative values indicate that the gap between the generic Democrat and the generic American is smaller than the gap between the generic Republican and the generic American, while positive values indicate the opposite.

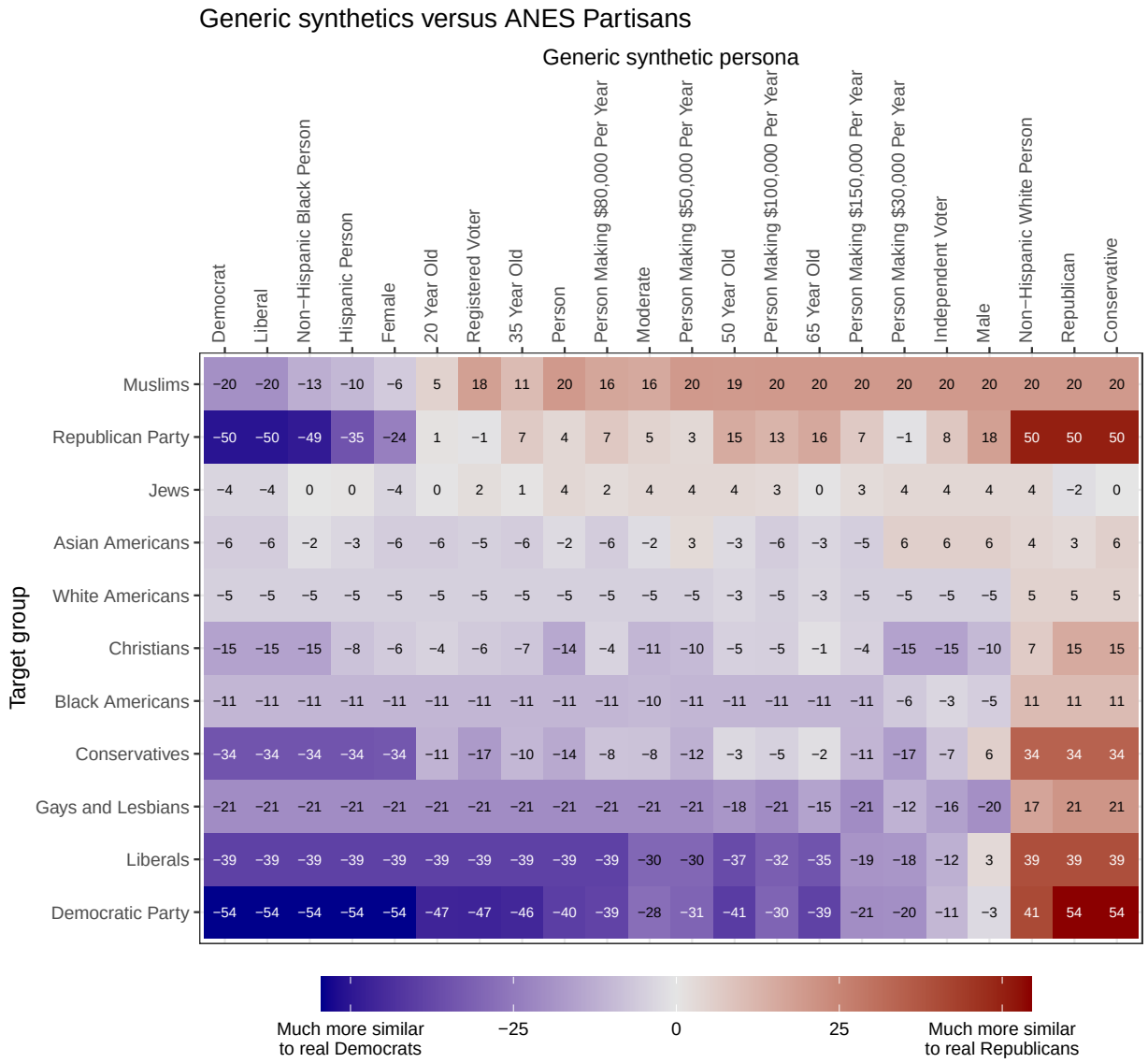

Figure 26: Similarity between generic synthetic ChatGPT respondents (y-axis) and human ANES Democrats (negative values in blue) and Republicans (positive values in red) across a range of target groups (y-axis). Majority of personae are more similar to human Democrats across a majority of target groups.

8 Time and Change

Our original results are based on a prompt that instructed the AI to role play as one of several different types of respondents in the year 2019. We chose this year for two reasons. First, it is temporally proximate to when the training data ends for ChatGPT 3.5 (September 2021). Second, it is prior to the Covid-19 pandemic, which we suspect would influence politically relevant beliefs – potentially including feeling thermometers. Combined, our goal was to find the easiest possible test case for the LLM.

However, our ANES data is not from 2019, but instead is two waves fielded in 2016 and 2020. Below, we test the sensitivity of our conclusions to the detailed prompt that explicitly links the human and synthetic respondents by year. We first examine whether the LLM performs worse when asked to provide responses from subjects in 2016 versus 2020. If ChatGPT's performance is temporally bound, it raises serious concerns about the use of synthetic samples generated by the LLM. Figure 27 visualizes the results of a simple analysis wherein we calculate the per-respondent mean absolute error and then average across groups and years. Black borders indicate a statistically significant difference, and dark bars represent the 2020 vintage. As illustrated, there is little evidence of significant performance loss in 2016 versus 2020. If anything, the LLM is worse at predicting attitudes towards Christians in 2020, and better at predicting attitudes towards Gays and Lesbians in 2020. Across all other groups, the differences are both statistically and substantively negligible.

We then test whether the empirical over-time change in group feeling thermometers found among the human subjects is replicated in the synthetic data. To test, we run a regression on the stacked dataset where we predict the feeling thermometer for each respondent as a function of the year interacted with data source and the party ID, implementing fixed effects for the target group. The predicted values of this regression are presented in Figure 28, illustrating little evidence of differences in the overtime change. However, when we run the same specification but include the interaction with the target group, we find divergent over time changes in attitudes towards white Americans in particular (see Figure 29. In general, the LLM data appears to consistently report the same or warmer attitudes among all partisans for all target groups in 2020 relative to 2016, while these patterns vary among the humans surveyed in the ANES.

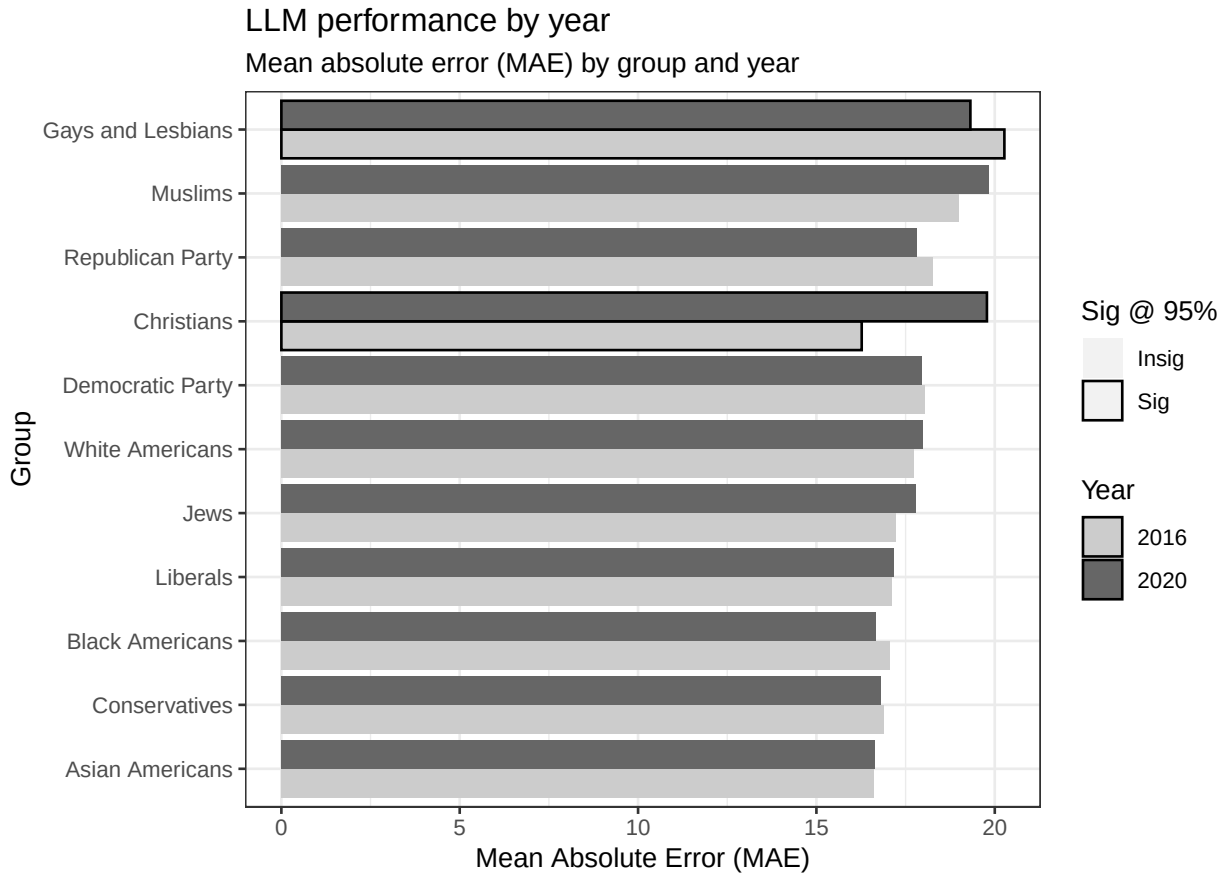

Figure 27: Mean absolute error (MAE, x-axis) measuring the absolute difference between the ANES feeling thermometer score for a given group (y-axis) averaged across all human respondents in 2016 (light gray bars) and 2020 (dark gray bars). Solid borders indicate differences between the 2016 and 2020 ANES waves where the MAE is significantly different.

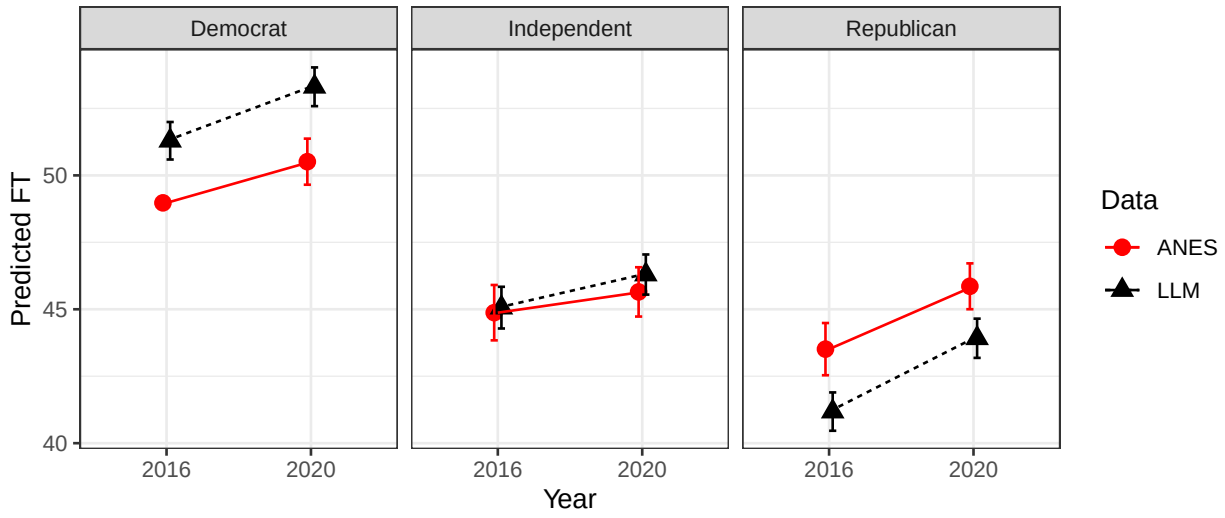

Figure 28: Predicted feeling thermometer scores (y-axis) estimated in the ANES (red points with solid lines) and synthetic data (black triangles with dashed lines) in the 2016 and 2020 waves (x-axes) by party ID (columns).

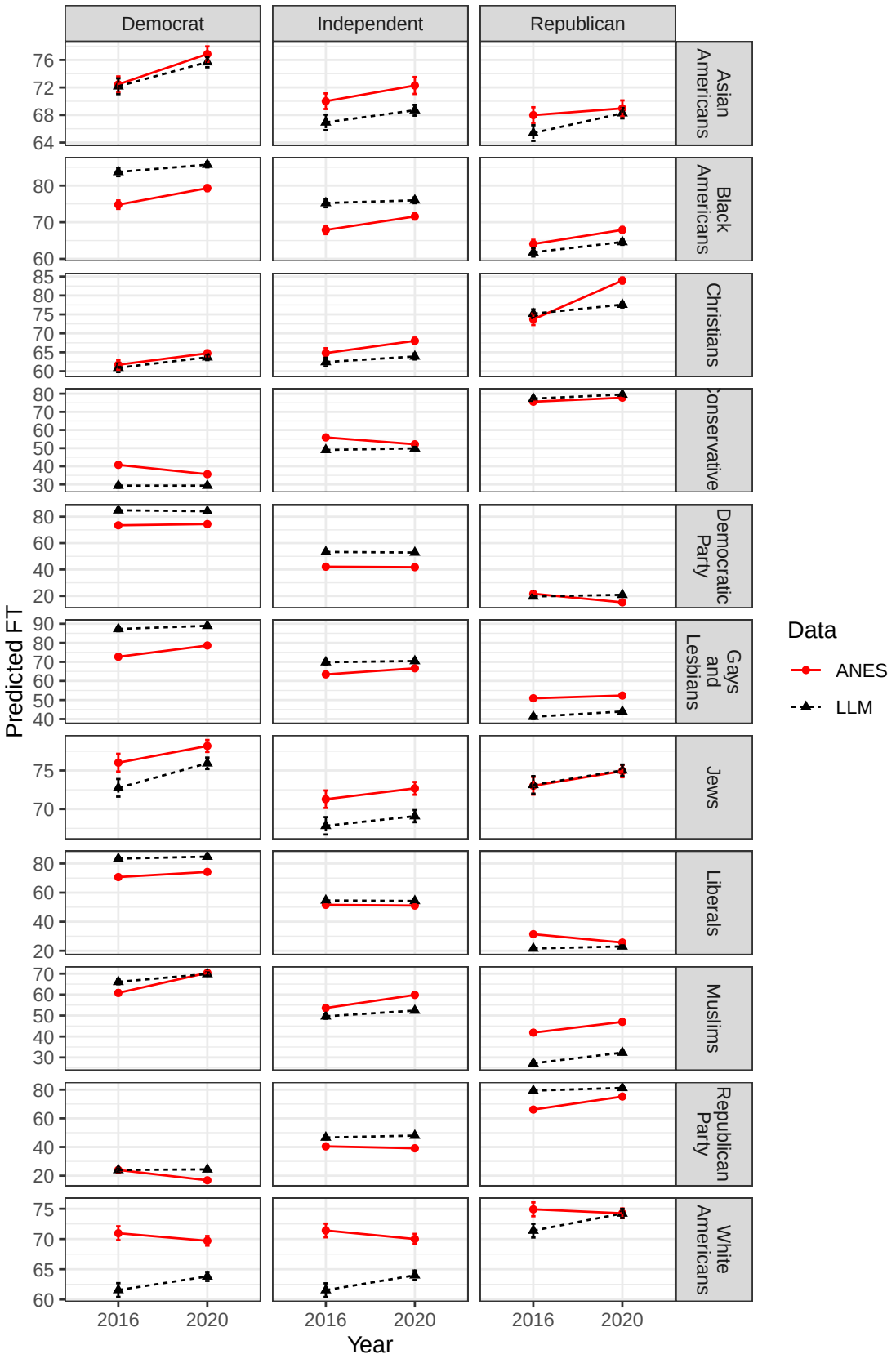

Figure 29: Predicted feeling thermometer scores (y-axis) estimated in the ANES (red points with solid lines) and synthetic data (black triangles with dashed lines) in the 2016 and 2020 waves (x-axes) by party ID (columns) and target group (rows).

9 Confidence, uncertainty and the empirical distribution of AI

ChatGPT is a generative LLM, meaning that it is optimizing to predict the most likely next word. The temperature hyperparameter in the AI governs how deterministic this process is. Higher temperatures mean that ChatGPT doesn't always choose the highest probability next word, but can instead choose the second highest, or third highest, and so on. In principle, this means that the results produced by our prompts are characterized by their own probability distribution. While our main findings concluded that the LLM's estimates were overly precise relative to real humans, it remains unclear what the actual nature of this posterior probability distribution is.

9.1 GPT's Self-Reported "Confidence"

Unfortunately, ChatGPT does not allow users to directly observe the posterior probabilities for each output.[4] However, we tried asking the AI to provide its confidence in the feeling thermometer score it provided. Our prompt yielded almost 900,000 responses that were non-numeric in nature, the vast majority of which adhered to a 5-item scale ranging from not at all confident to very confident, although the precise wording varied. In addition, we note that there were very few examples of responses that indicated lower levels of confidence (i.e., "somewhat", "moderately", "medium"). We plot the distribution of these truncated responses for reference in Figure 30, but drop them from our main analyses that follow which focus only on responses that included a number between 0 and 100 percent.

Turning to the numeric responses provided by the AI, we again see that they are mainly confident, with the majority of values above 80 (see Figure 31).

We can plot these measures of the AI's subjective confidence against two dimensions of variation where we might expect to find less confidence. The first is in the richness of the prompt description. As illustrated in Figure 32, we show that the average self-reported confidence for responses provided in the demographics-only prompt are lower than those provided in the politics-only or combined prompts, with the latter recording the highest overall average. Substantively, this seems sensible: when the AI is given a vaguer sketch of the persona they are meant to adopt, its confidence in the responses declines. Interestingly, this is most pronounced for the feeling thermometers about politically salient groups (Democratic Party, Republican Party, liberals, and conservatives). Without knowledge of the persona's political identity, the LLM is less confident in its guess at the persona's feeling thermometer toward political groups.

The second dimension is the feeling thermometer itself. All else equal, we might expect that a neutral feeling thermometer is associated with weaker confidence, whereas either very cold or very warm feelings are associated with greater confidence. This expectation is again supported in the data, as visualized in Figure 33. Where we have sufficient coverage across the support of feeling thermometer scores (x-axes) we see a U-shaped pattern in the LLM's confidence.

The preceding analyses rely on the LLM's self-reported confidence to understand the variability of the synthetic data it generates. However, as we show in Figure 34, the relationship between self-reported confidence (x-axes) and empirical uncertainty (standard deviation of feeling thermometer scores associated with a certain group and confidence, y-axes) is – if anything – positive. This means that the more confident is ChatGPT in its feeling thermometer score for a given group, the more variation we observe in the scores. Note that this plot is generated by calculating the average confidence for all thirty synthetic draws for each human respondent for each target group, along with the standard deviation of the generated feeling thermometer scores, meaning that these patterns

[4] Other tools such as Da Vinci, OpenAI's LLM-based classifier, do provide such information.

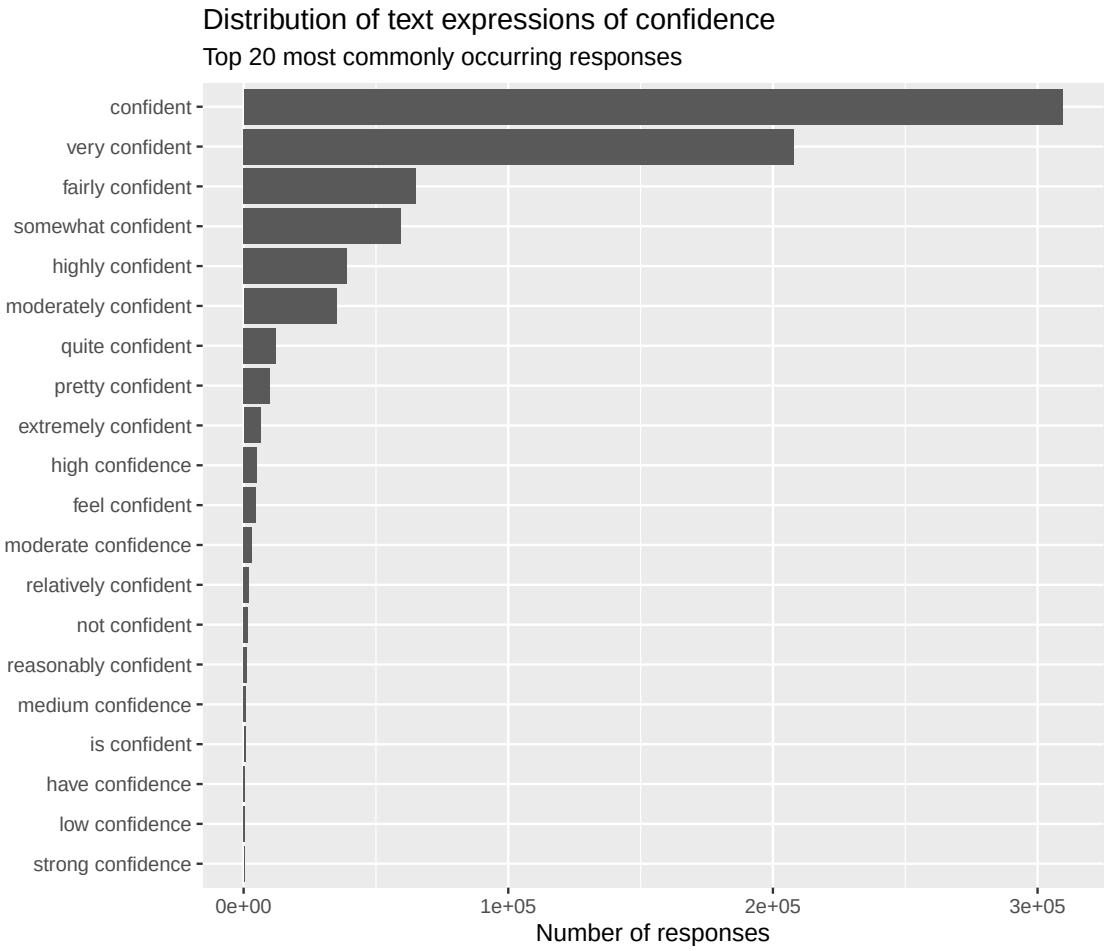

Figure 30: Descriptive prevalence of non-numeric confidence responses, truncated to only include the word stem "confid" and the word prior.

are not driven by heterogeneity across partisanship or some other attribute. While some of the target groups exhibit less of a strong positive association than others, we nevertheless underscore that none of the evidence suggests that self-reported confidence is a reasonable proxy for empirical uncertainty.

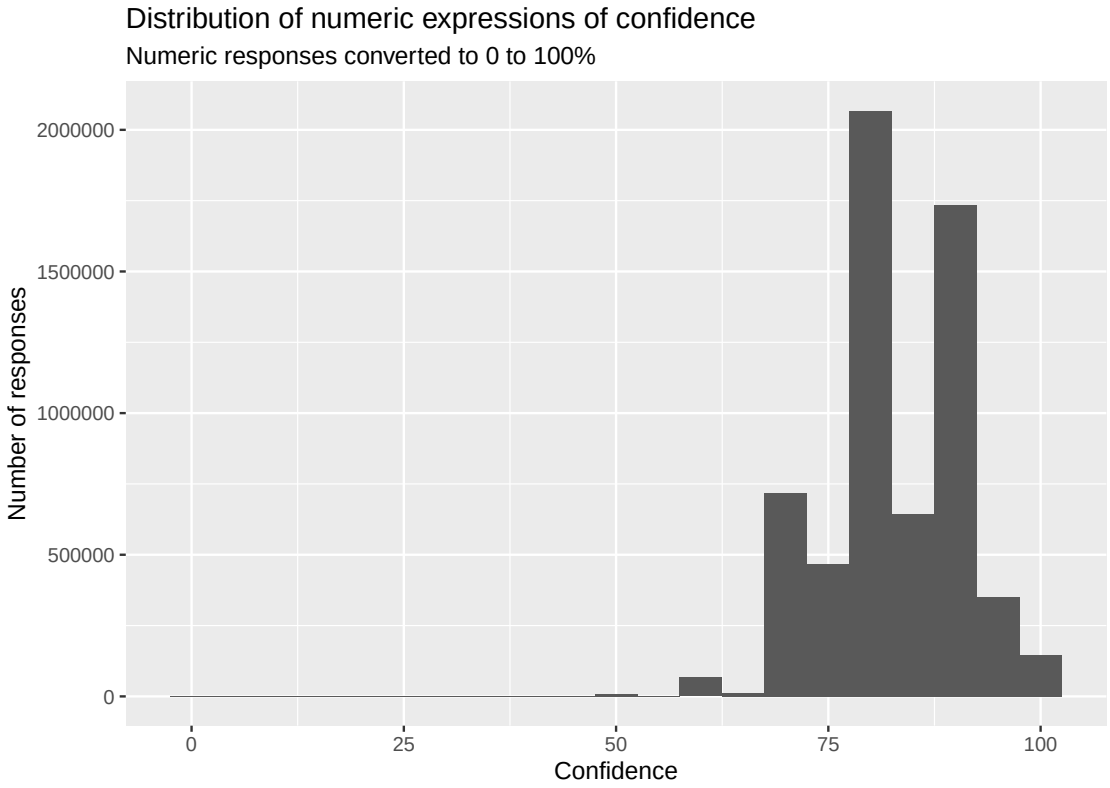

Figure 31: Empirical distribution of numeric confidence responses.

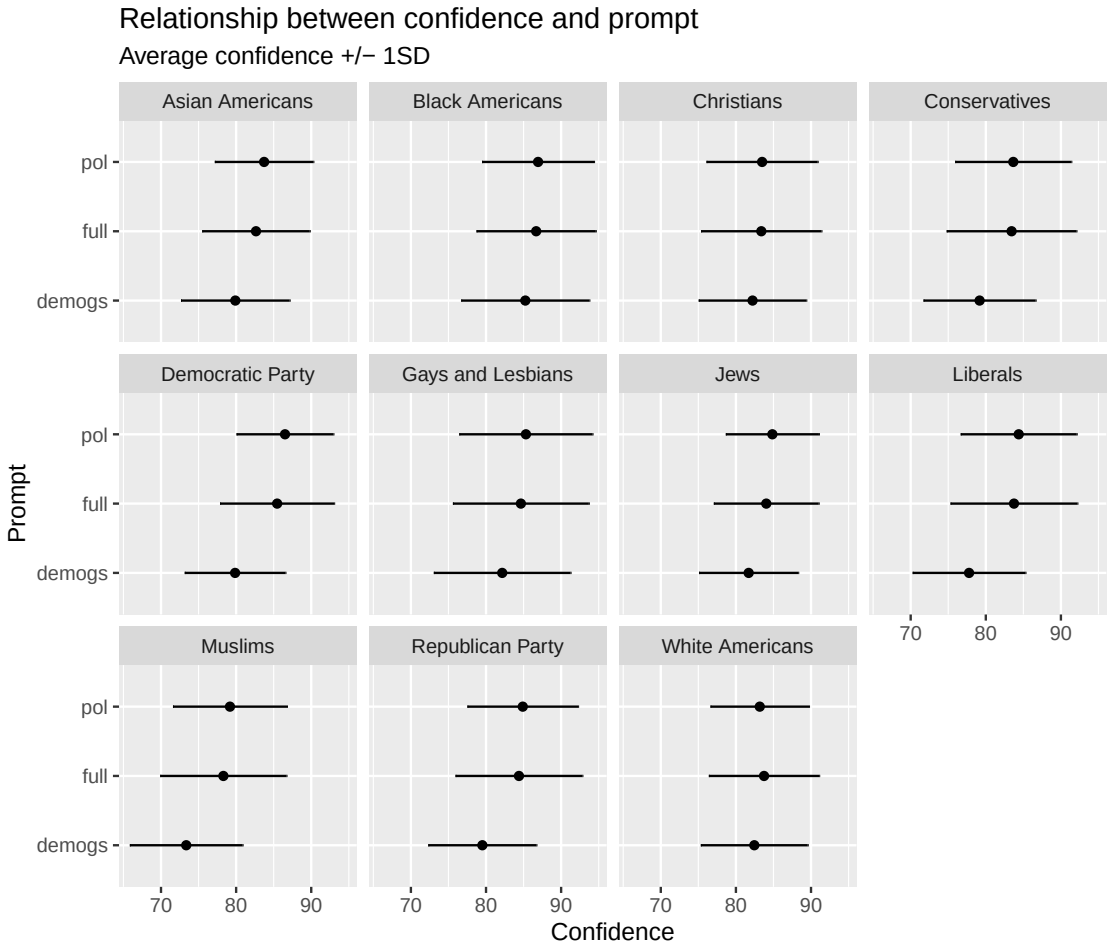

Figure 32: Average confidence (points) plus / minus one standard deviation (bars) indicated on the x-axes, as a function of the prompt (y-axes) and target group (facets).

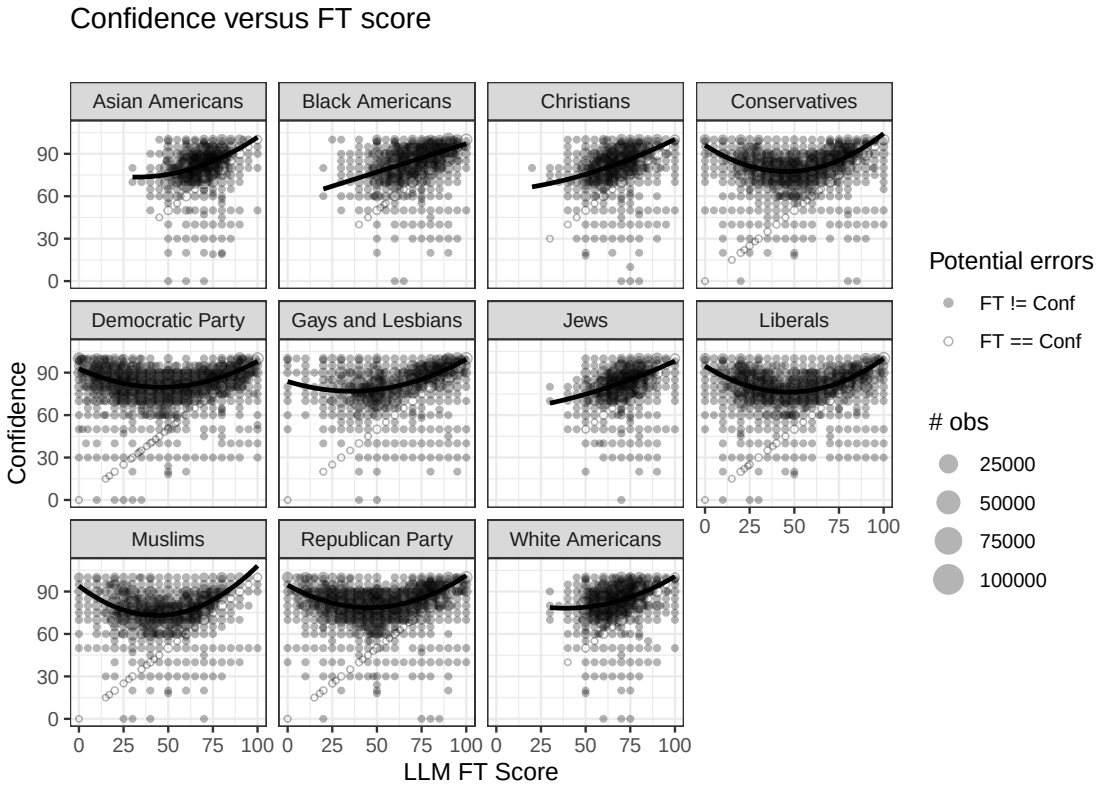

Figure 33: Feeling thermometer score produced by ChatGPT (x-axes) against numeric confidence associated with the score (y-axes) by target group (facets). Points where the feeling thermometer score exactly equals the confidence highlighted with hollow white points, potentially reflecting errors in the LLM's response.

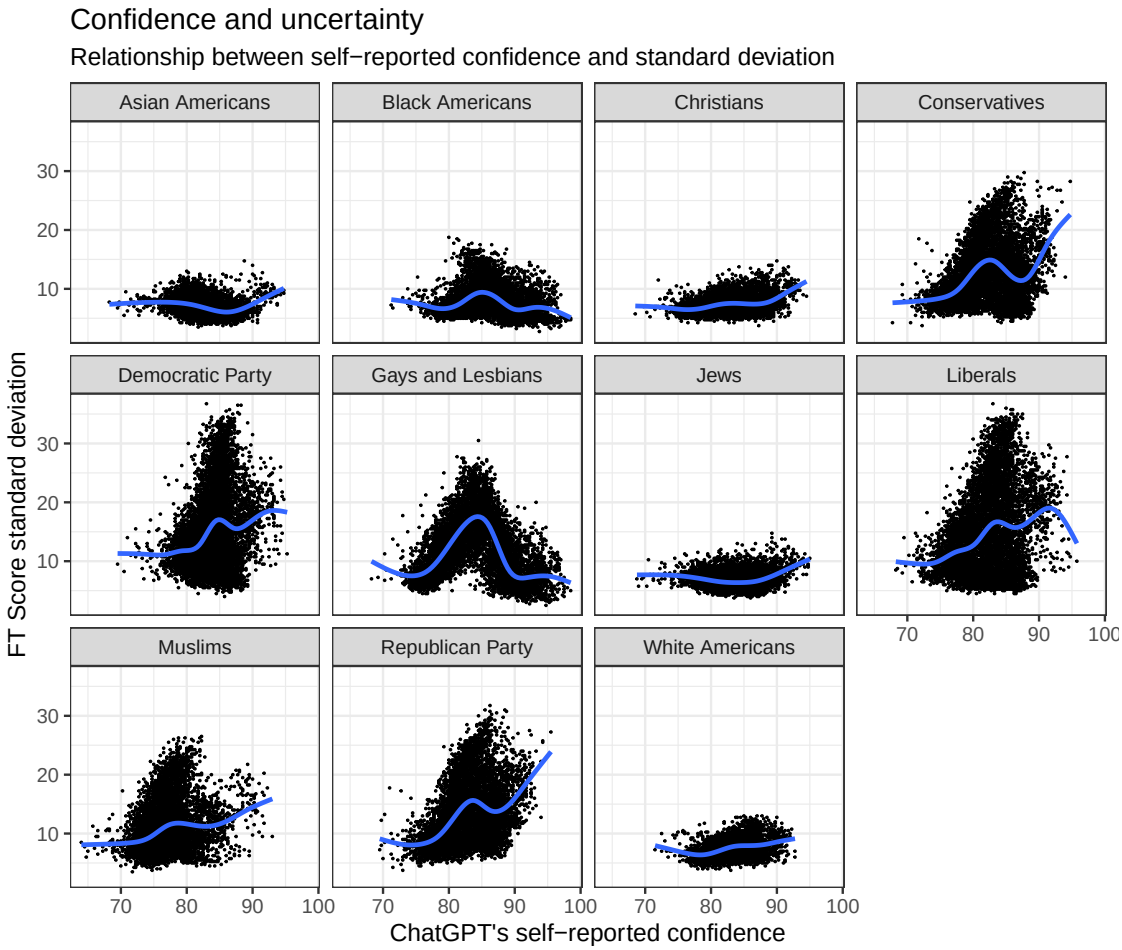

Figure 34: Y-axes indicate the standard deviation of 30 synthetic samples for each human respondent-by-target group (columns)-by-prompt type. X-axes indicate the average numeric confidence associated with these 30 samples.

9.2 Empirical Uncertainty

How then can we understand the uncertainty of the AI? As discussed in the main text, ChatGPT produces far less variation in its synthetic data, relative to real humans. Would this change if we provided a more detailed description of each human respondent? In theory, if the algorithm was given a rich characterization for each person, it might do a better job predicting their actual attitudes towards different groups, and thus better reproduce the empirical variability of these responses when aggregated for statistical analysis. Yet as illustrated in Figure 35, the per-human variation doesn't actually decline with richer descriptions. As above, we generate this plot by calculating ChatGPT's standard deviation for the synthetic approximations of each human respondent's feelings by group and by prompt. While the LLM's variability is largest in the demographics-only condition, it is smallest *not* in the combined prompt, but rather in the politics-only prompt.

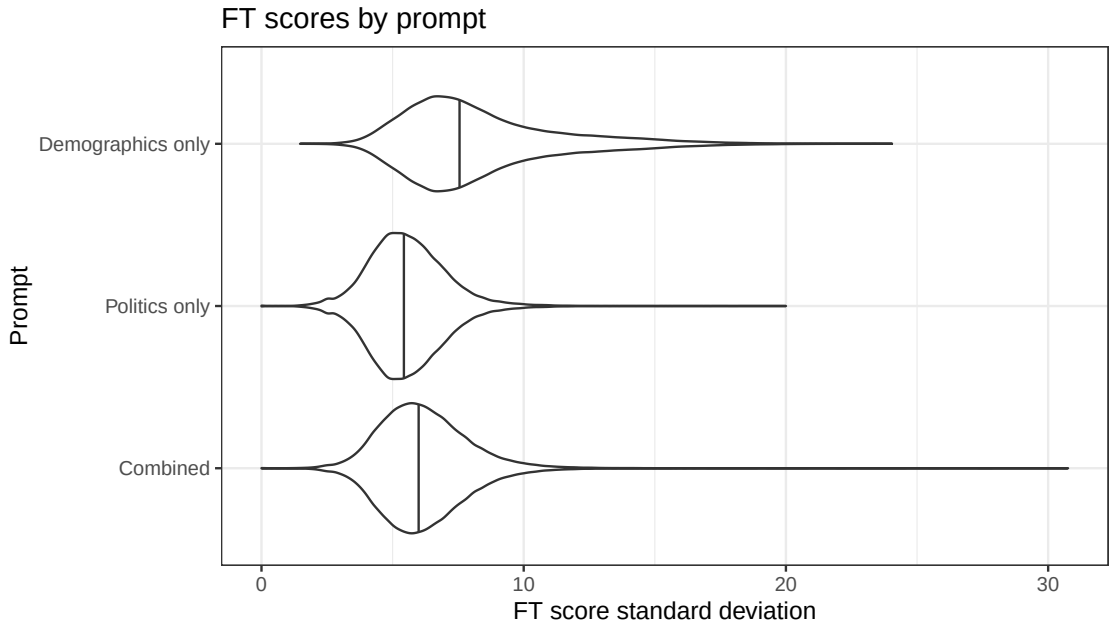

Figure 35: Distribution of standard deviations (x-axis) calculated across 30 synthetic samples for each human respondent-by-target group-by-prompt type (y-axis).

One possible explanation might be that less frequently occurring groups are more difficult for the algorithm to predict since they constitute a smaller part of the training data. With 30 samples per human-by-prompt-by-target group, we start with a measure of the standard deviation at this unit of analysis and plot the average standard deviation by the number of humans who make up each profile. As illustrated in Figure 36, there is little evidence that ChatGPT is more confident for groups that are more commonly occurring in the data, contrary to expectations.

It may be that ChatGPT struggles more with contradictory descriptions, such as a liberal Republican or a conservative Democrat. We do find some evidence for this expectation in Figure 37, which shows marginally larger standard deviations for more extreme ideological groups, especially where the ideology is misaligned with the party. The average synthetic standard deviation for the most commonly-appearing groups is slightly less than 7 feeling thermometer degrees, while the same measure for the least commonly-appearing groups is more than 10.

Yet even in these instances where we record larger measures of uncertainty for the AI, we nevertheless underscore that these are far more precise than what we observe among real humans.

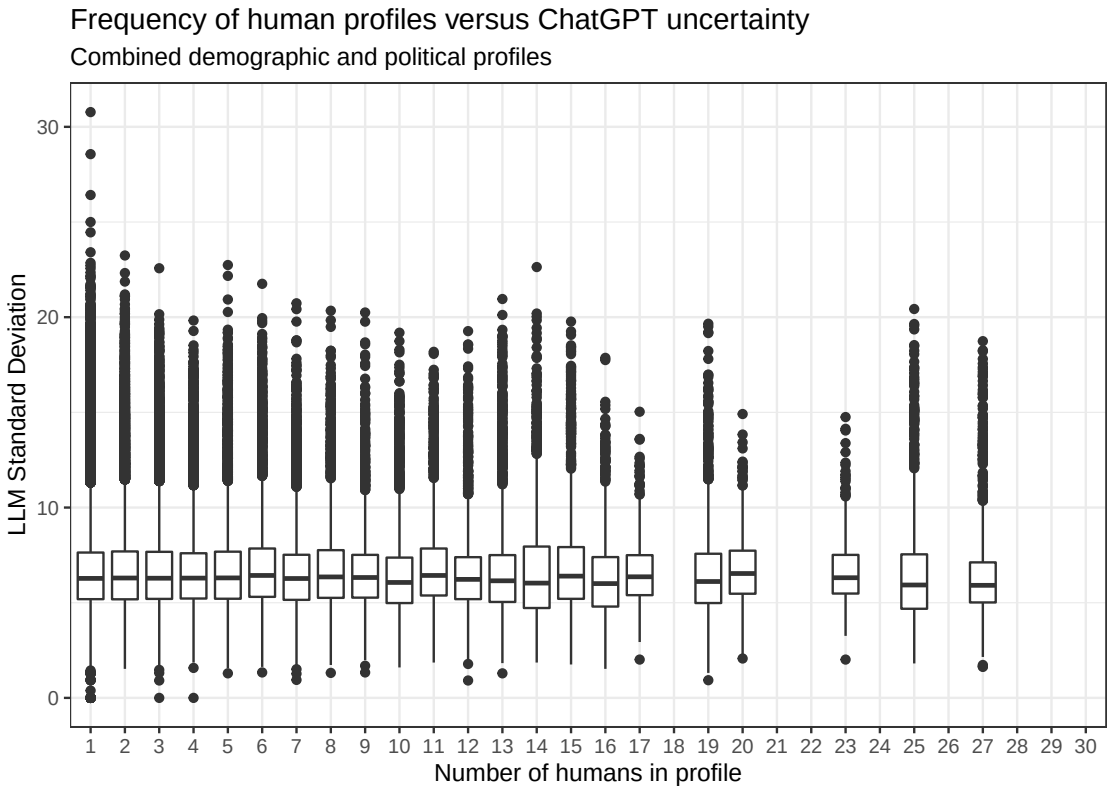

Figure 36: Distributions of standard deviations (y-axis) calculated across 30 synthetic samples for each human respondent-by-target group in the most detailed prompt by the number of human respondents associated with the persona's profile described in the prompt (x-axis).

First, we reproduce the Figure 37 below but include the standard deviation of the ANES measures for contrast (Figure 38). The figure illustrates that the curvelinear pattern documented in the synthetic data is actually an accurate reflection of the uncertainty among human respondents, despite being far less noisy. Extreme ideologues appear to be more volatile in both instances.

Second, we evaluate whether these differences in uncertainty also are reflected in the overall performance gap between the ANES target and the synthetic data. Figure 39 compares the mean absolute error (MAE, top row) and the ratio of standard deviations (bottom row) across the same bins, confirming the conclusion that the synthetic data is further from the ANES among profiles that are less commonly occurring, although interestingly the gap in the variation narrows here, likely due to the inflated standard deviations found in the synthetic data (see Figure 37).

One final question might be whether these patterns reflect solely the sample sizes or if they are also picking up the paradox of an extremely liberal Republican or an extremely conservative Democrat. Presumably among the few humans who fit these descriptions, variation in the ANES FT scores is driven by genuine variation in their feelings towards these groups. Does the variation in the synthetic data also reflect this natural variation? Or is it instead a product of the LLM sometimes prioritizing the ideological description and in other cases prioritizing the partisan description? (And if it is the latter, is this substantively different from what we might imagine is going on in the minds of those humans who self-identify in these non-aligned cells?)

Importantly, we find less evidence of these relationships when we disaggregate further, suggesting that the challenge is less to do with sparsely populated profiles per se, and more to do with

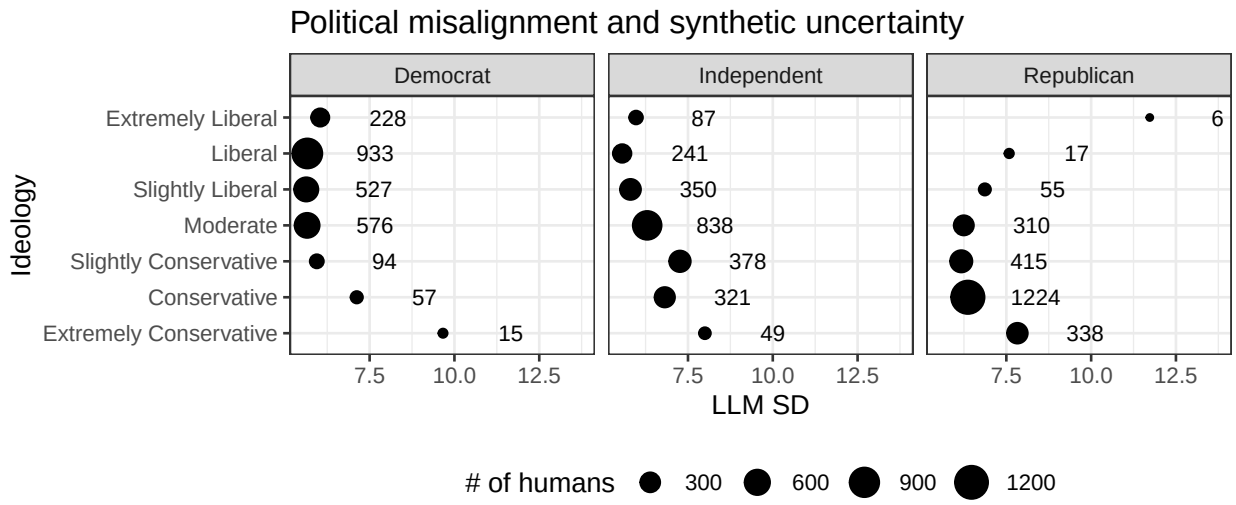

Figure 37: Average of standard deviations (x-axes) calculated across 30 synthetic samples for each human respondent-by-target group in the most detailed prompt by ideology (y-axis) and party ID (columns) of the persona described in the prompt. Points are sized and labeled with the total number of human respondents associated with each ideology-by-partisanship bin.

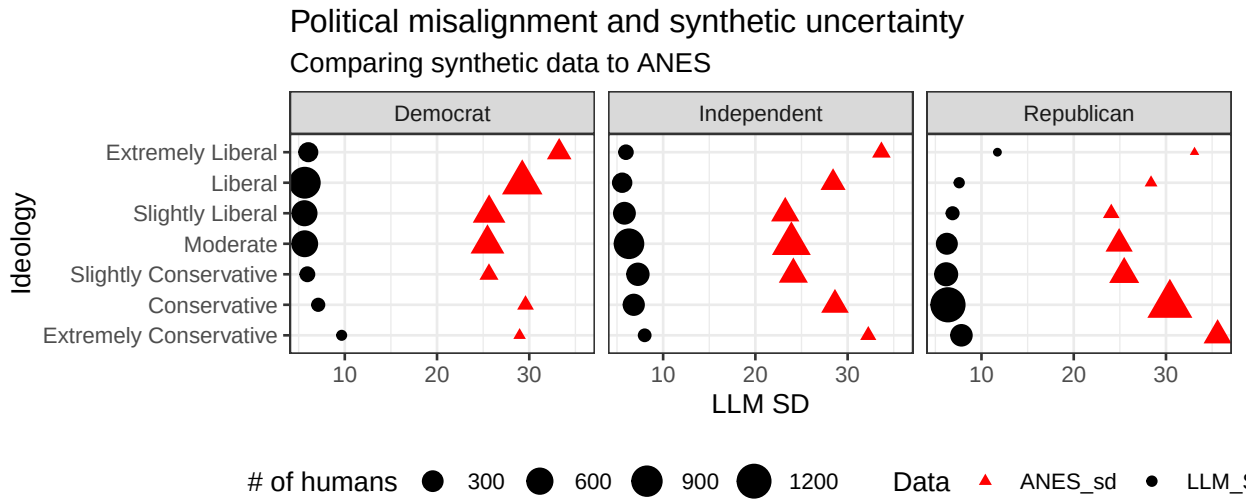

Figure 38: Average of standard deviations (x-axes) calculated across 30 synthetic samples for each human respondent-by-target group in the most detailed prompt by ideology (y-axis) and party ID (columns) of the persona described in the prompt indicated with black points. ANES standard deviations across all humans in an indicated ideological-by-party ID bin indicated with red triangles.

misaligned political attributes. Figures 41 and 40 plot ChatGPT's standard deviation (x-axes) for all covariate profiles for which we have at least 10 human respondents, and includes the standard deviation associated with the ANES data for comparison. For almost every profile across every target group, we see that the LLM's estimates are far more precise than those associated with real humans, especially as we look among profiles with larger samples.

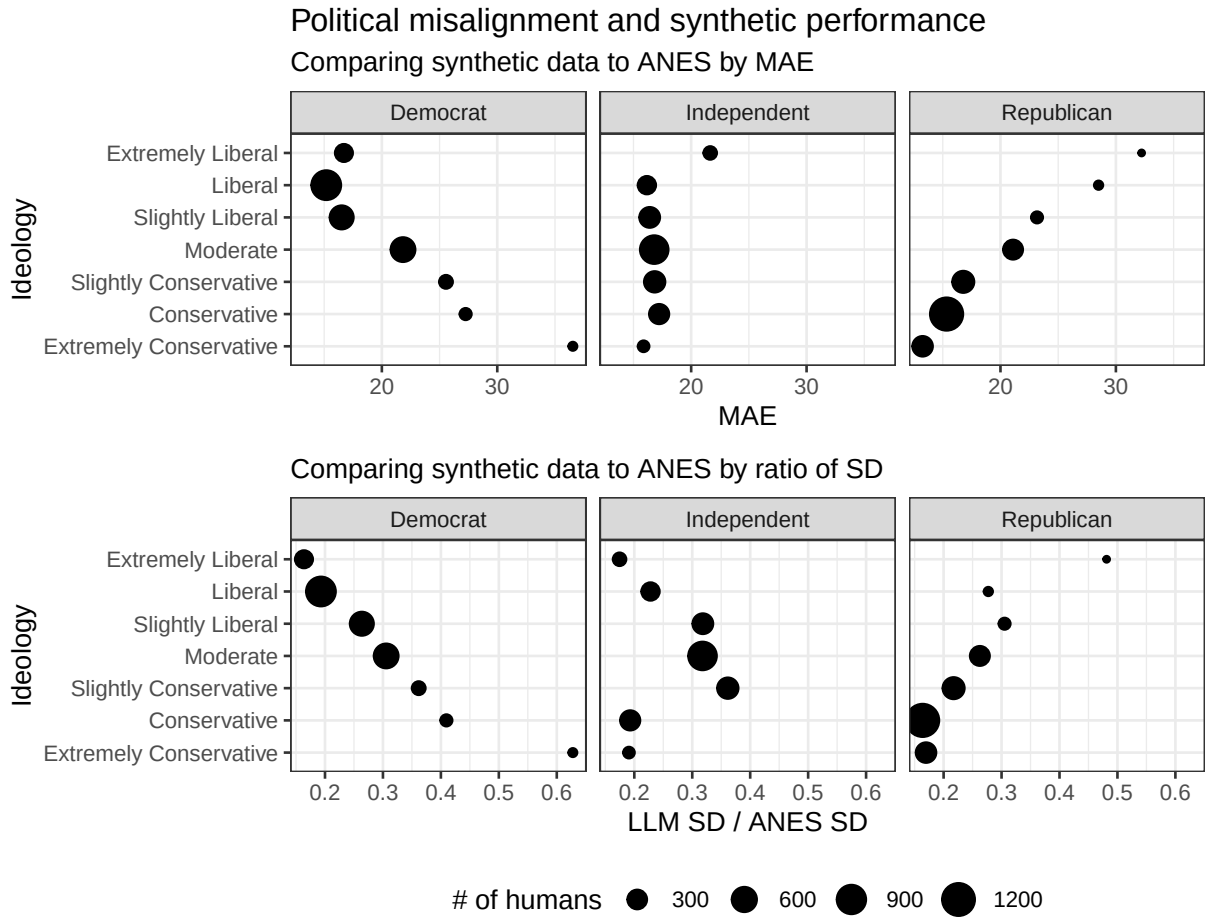

Figure 39: Evaluations of how similar the synthetic data is to its human counterparts in the ANES. Top three plots summarize difference in terms of mean absolute error (MAE) by ideology (y-axis) and partisanship (columns). Bottom three plots summarize differences in terms of the ratio of the synthetic standard deviation (calculated across 30 synthetic samples for each human respondent-by-target group in the most detailed prompt) over the standard deviation of all humans falling into one of the ideology-by-partisanship bins.

9.3 Posterior Distributions

A final point of interest is what the empirical posterior distributions actually look like. With 30 synthetic respondents per human, we aggregate over all covariates included in the full prompt and plot the empirical distributions by profile and target group. Figure 42 plots the distribution for synthetic feelings toward the Democratic Party by all personas (of which there are 7,530 associated with human respondents). Horizontal bars represent the interdecile range and vertical marks capture the median, while colors reflect the partisan affiliation of the persona. Empirical distributions are ordered first by the median, then by the upper bound of the interdecile range.

As illustrated, there is clear evidence of anchoring of the synthetic data to 5 unit increments, with the interdecile range exhibiting consistent terminations at units of 5. In addition, the plot reveals the highly consistent estimates of variance across all prompted personas, with the interdecile ranges spanning between 10 and 20 thermometer "degrees", and rarely spanning more than 20.

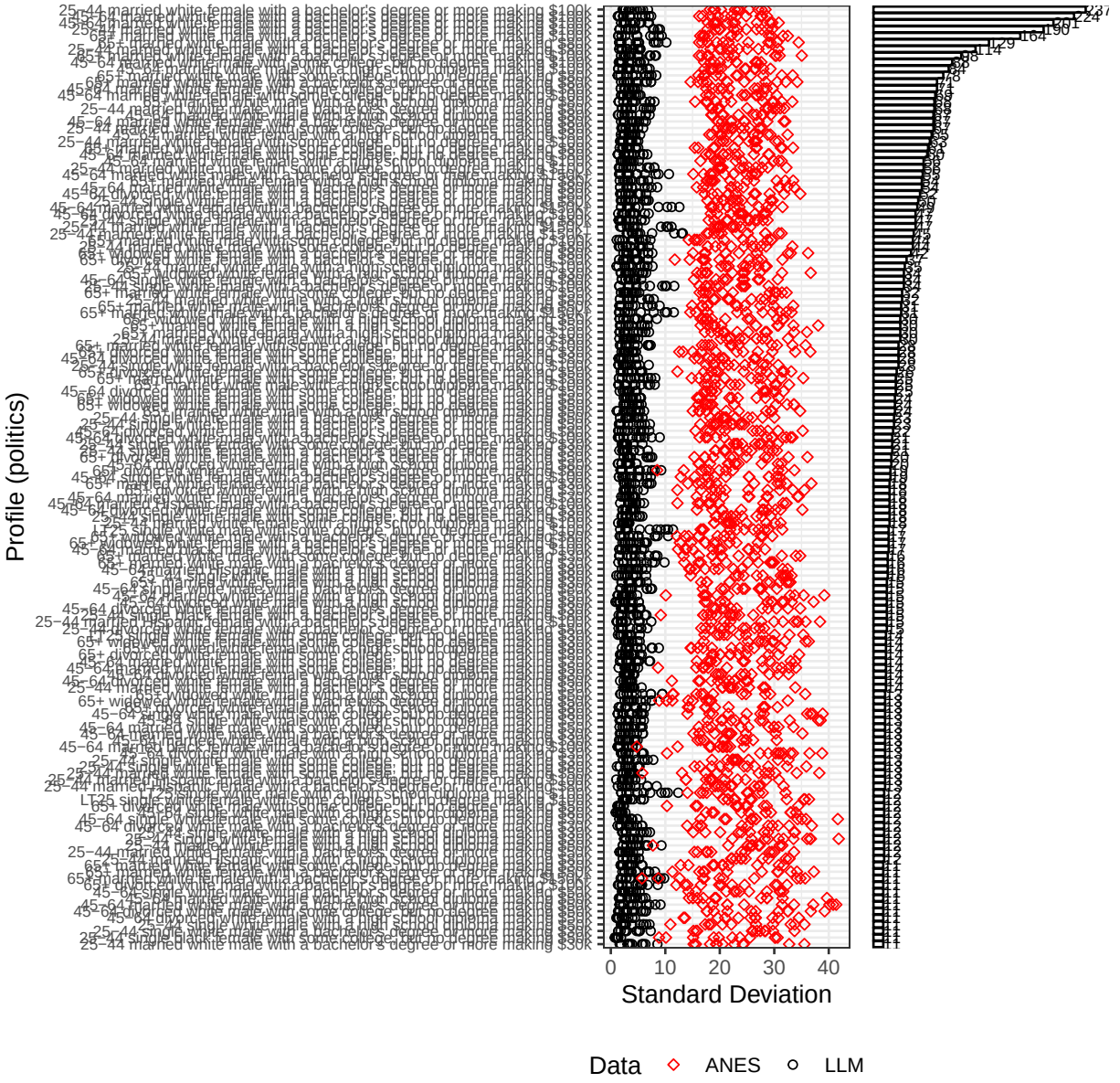

Figure 40: Standard deviation comparison between humans associated with a particular demographic profile (age by marital status by race by gender by education by income, y-axis) indicated in red diamonds, compared to the standard deviation calculated among their synthetic counterparts indicated in black circles for the demographic profiles with more than 10 humans in them (total counts indicated by bars on right).

Furthermore, most of the distributions appear reasonably symmetric, with the medians appearing roughly in the center of the interdecile range. However, in some cases, the empirical distribution exhibits substantial skew, which we highlight in Figure 43. While we find expected patterns between the direction of skew and political identities (reflecting the bottom and top censoring of the feeling thermometer distribution), we nevertheless also note that some of these instances do not reflect truncated distributions.

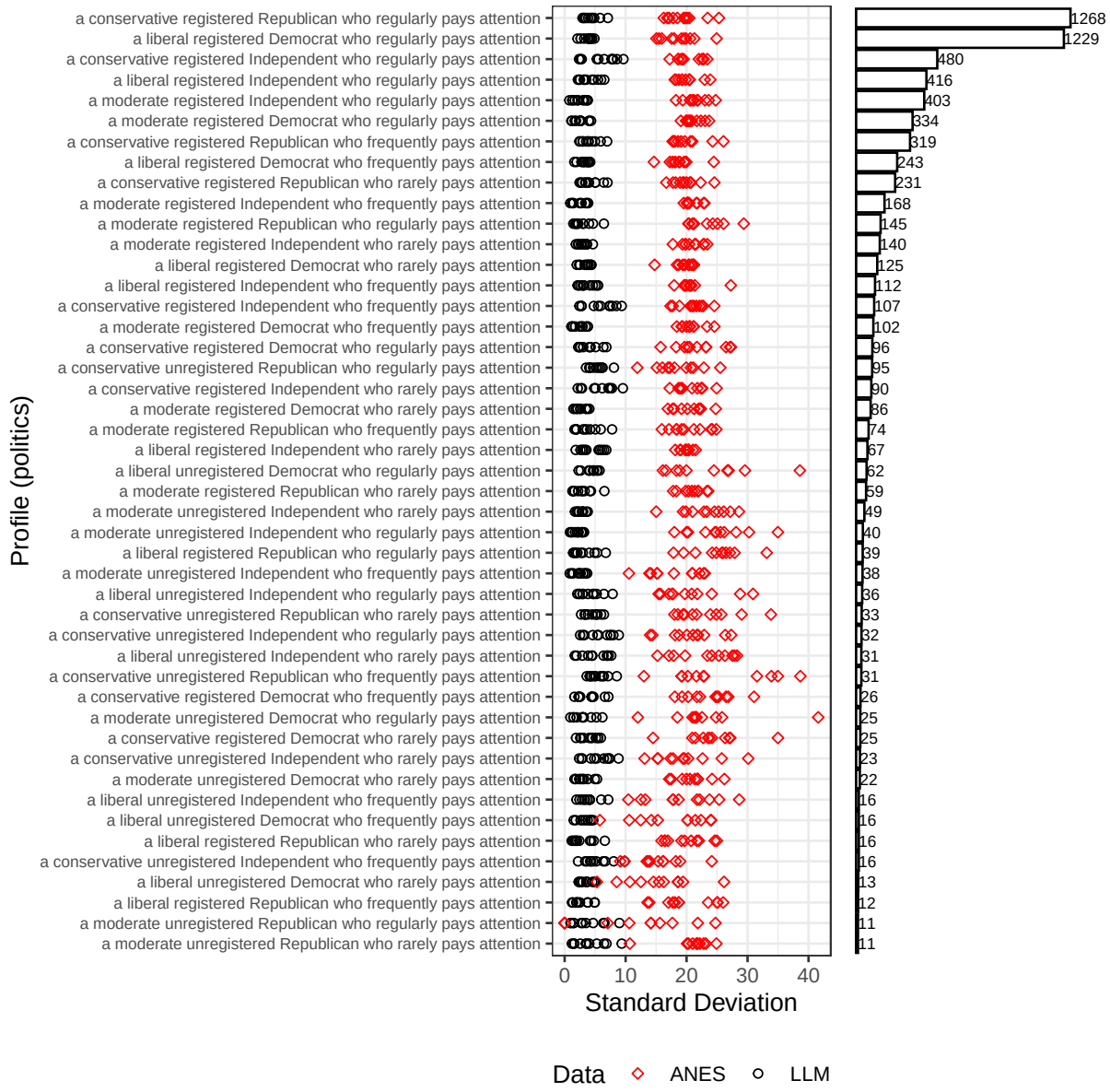

Figure 41: Standard deviation comparison between humans associated with a particular political profile (ideology by registration status by party ID by interest in news and politics, y-axis) indicated in red diamonds, compared to the standard deviation calculated among their synthetic counterparts indicated in black circles for the political profiles with more than 10 humans in them (total counts indicated by bars on right).

We illustrate similar patterns across other target group thermometers in Figure 44. Although the overall support for feelings towards racial groups are generally much narrower, never falling below 50, the per-persona interdecile range remains similar to those documented above – between 10 and 20 "degrees".

We end with some descriptive plots of the average per-persona standard deviation, broken out by target group (Figure 45) and by covariates (Figure 46). Despite there being some variation

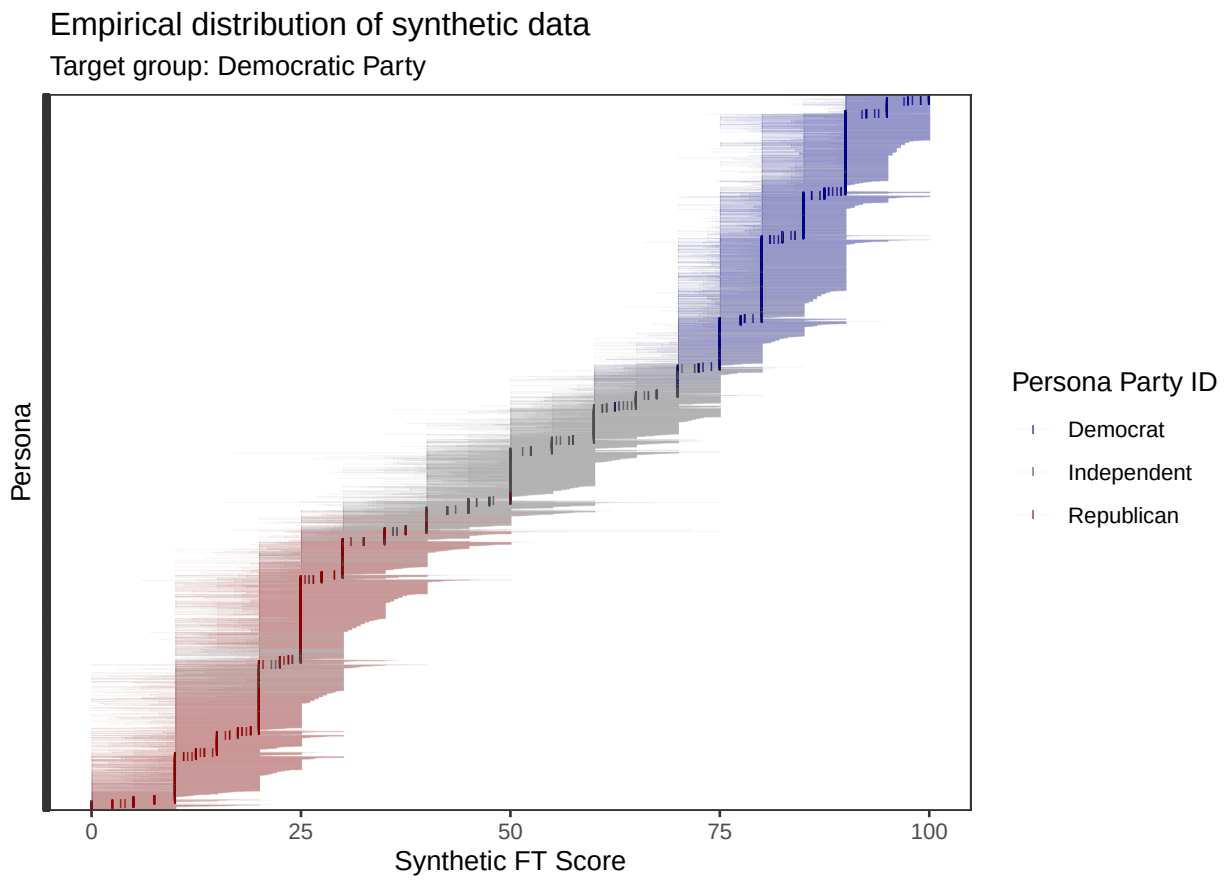

Figure 42: Median synthetic feeling thermometer score (vertical bars, x-axis) by all personas (ordered by thermometer score toward the Democrat Party, y-axis), along with interdecile range (bars) by the partisanship of the prompt (colors).

across target groups in terms of the synthetic data's variability (notably for the political parties), there is never a target group for which the synthetic data is as noisy as the ANES data.

And while there is also consistent evidence that the synthetic data estimates the attitudes of liberals and Democrats with less uncertainty than it does for conservatives and Republicans, these differences are no greater than 1.5 thermometer degrees. Notably, these differences pale in comparison to the differences in precision between the LLM and ANES data (see Figure 47).

Finally, Figure 48 plots the empirical distributions for every profile and target group, broken out by ideology and party affiliation.

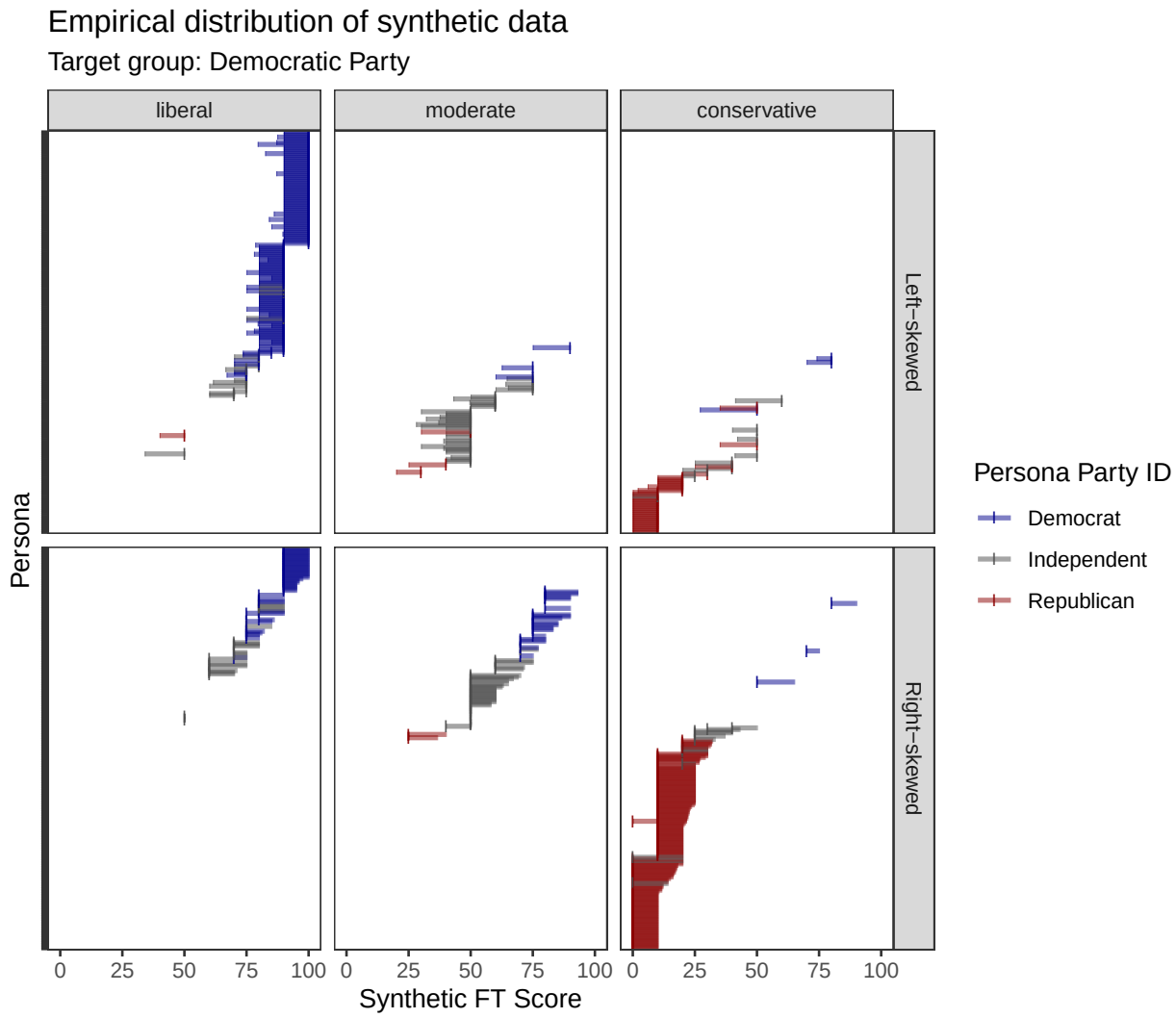

Figure 43: Median synthetic feeling thermometer score (vertical bars, x-axis) by all personas (ordered by thermometer score toward the Democrat Party, y-axis), along with interdecile range (bars) by the partisanship of the prompt (colors). Plots subset to profiles that are strongly skewed (rows) and coarsened ideology of prompt (columns).

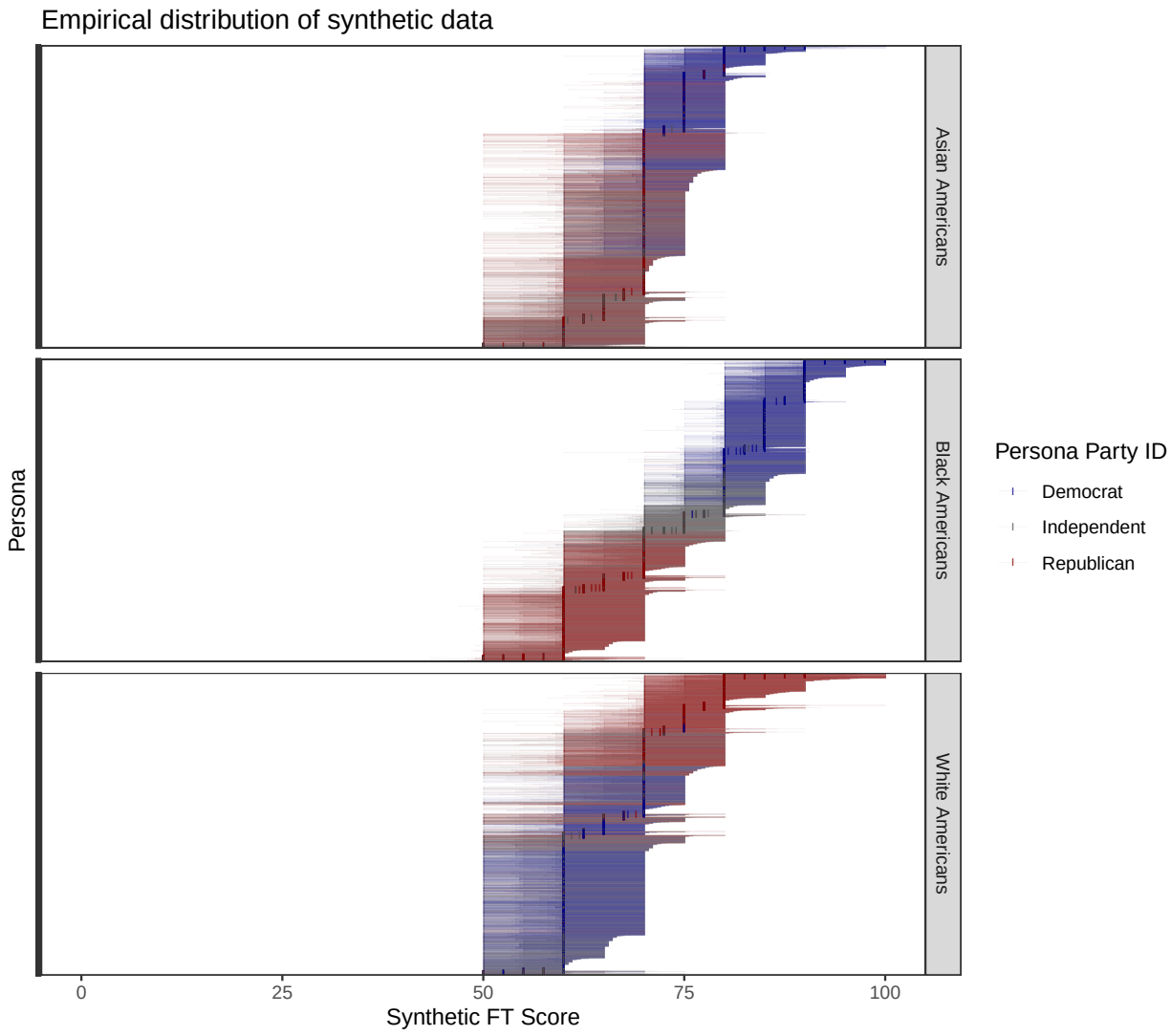

Figure 44: Median synthetic feeling thermometer score (vertical bars, x-axis) by all personas (y-axis), along with interdecile range (bars) by the partisanship of the prompt (colors), by race of target group (rows).

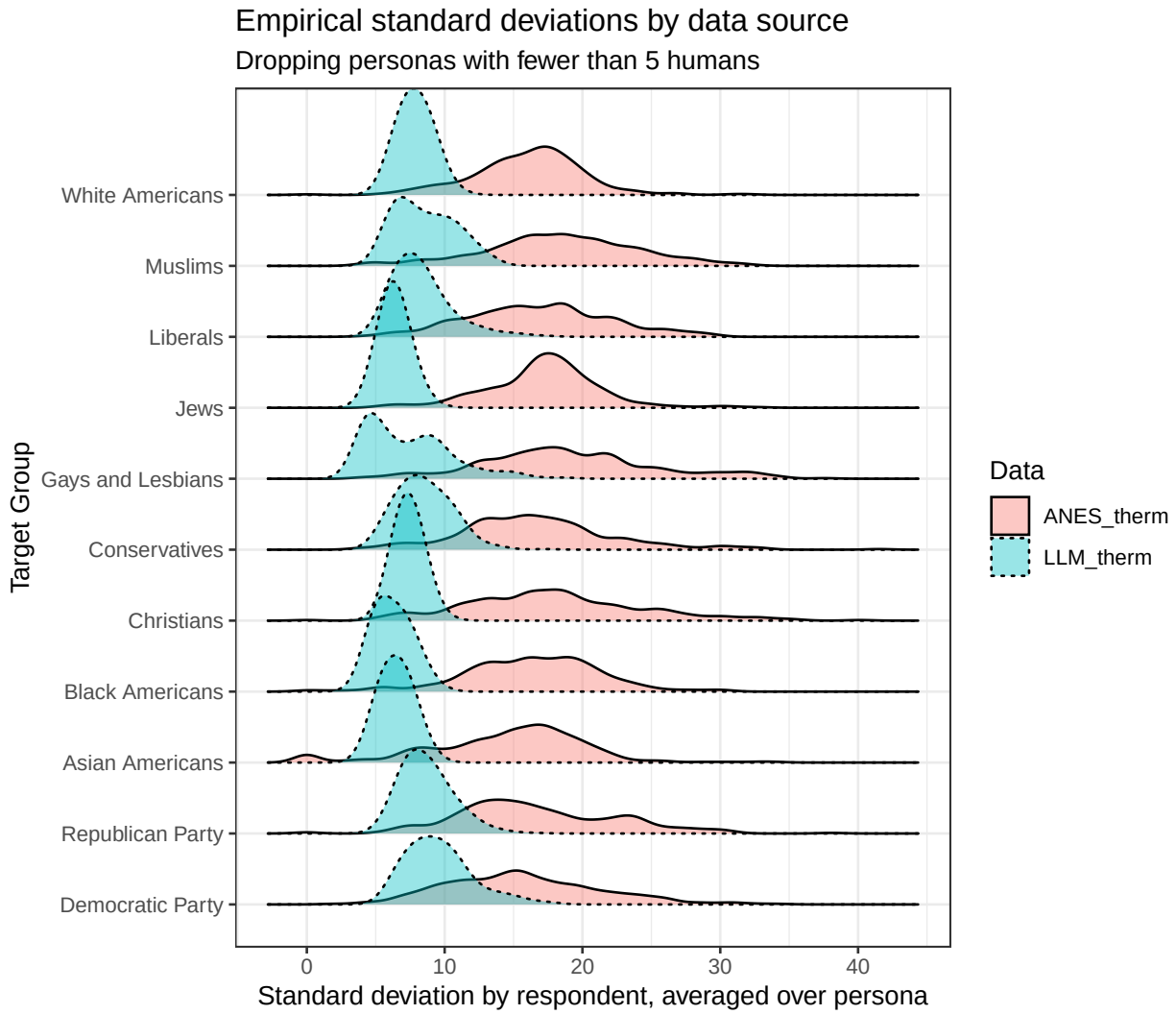

Figure 45: LLM standard deviation calculated across 30 synthetic samples for each human respondent-by-target group in the most detailed prompt (x-axis, teal distributions outlined with dotted lines) compared to ANES standard deviation calculated across all respondents in a given persona (red distributions outlined with solid lines) by target group (y-axis), dropping personas with fewer than 5 human respondents.

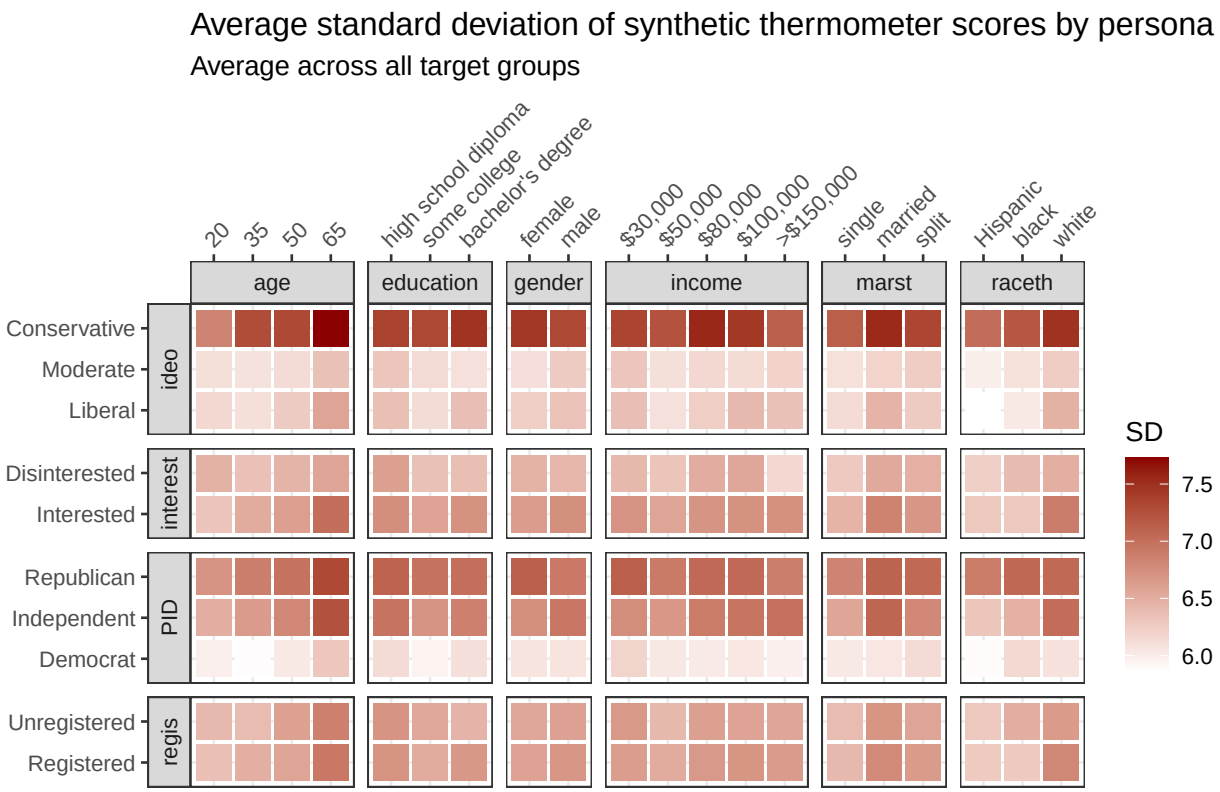

Figure 46: Average standard deviation in synthetic data, calculated across 30 synthetic samples for each human respondent-by-target group in the most detailed prompt, broken out by political covariates (y-axis) and demographic covariates (x-axis).

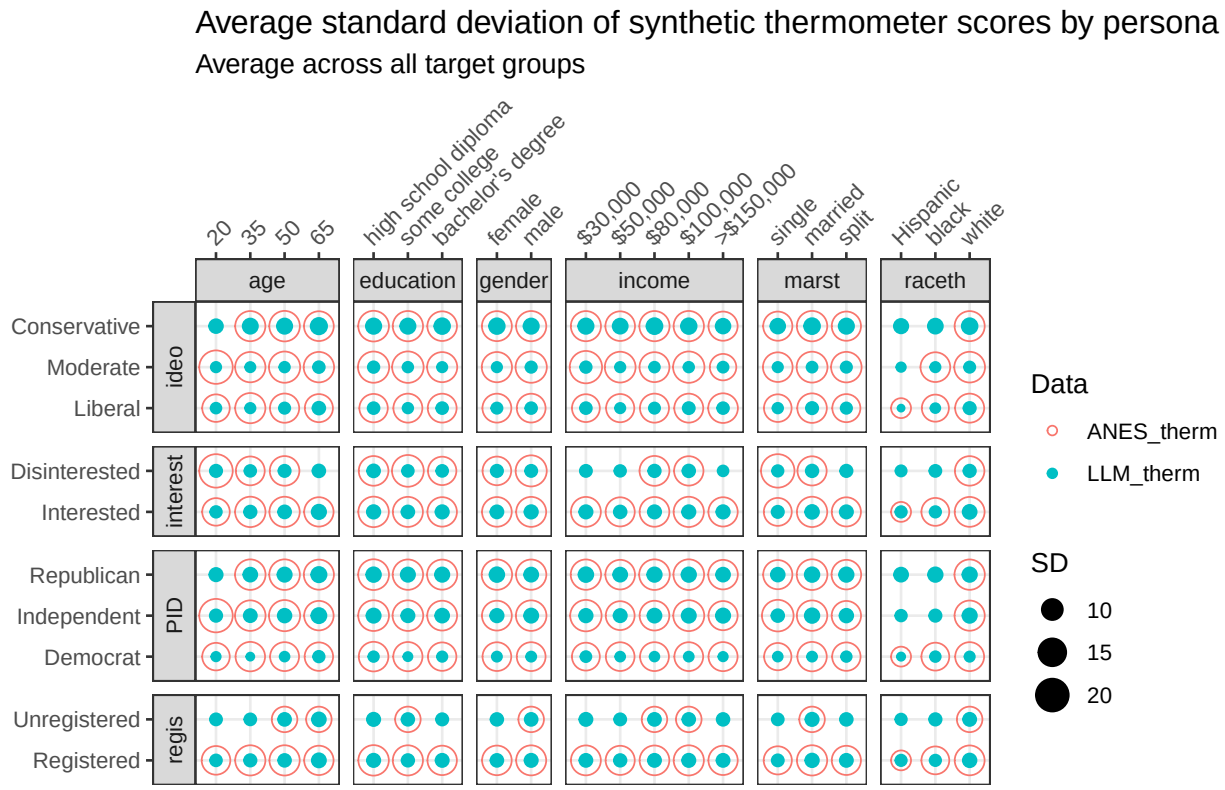

Figure 47: Average standard deviation in synthetic data (solid teal circles), calculated across 30 synthetic samples for each human respondent-by-target group in the most detailed prompt, broken out by political covariates (y-axis) and demographic covariates (x-axis). Hollow red circles indicate standard deviation for same political-demographic profile calculated in the ANES, restricting attention to personas with more than 5 human respondents.

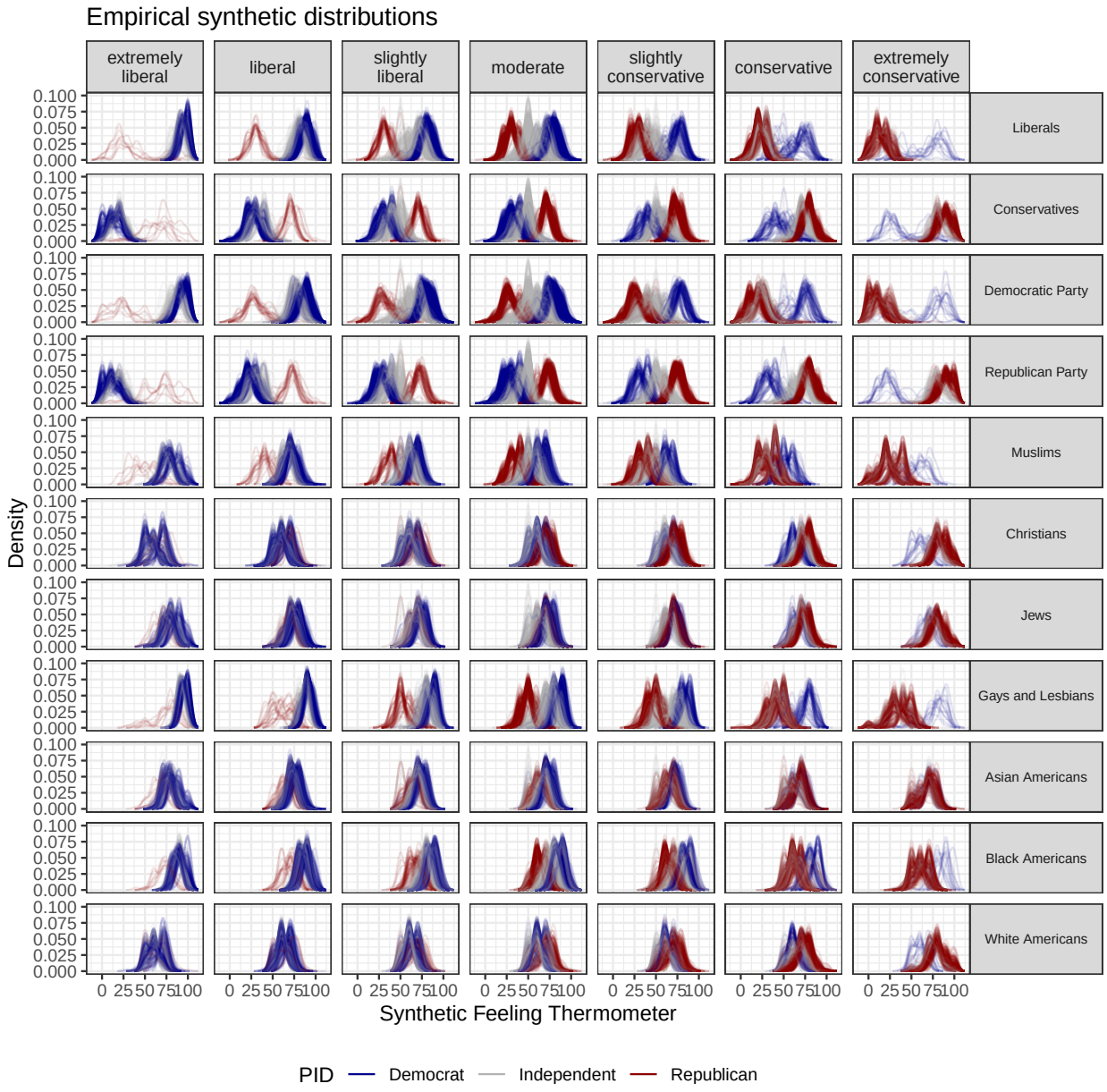

Figure 48: Empirical distribution of synthetic feeling thermometer scores in the most detailed prompt, broken out by ideology of persona (columns), partisanship of persona (colors) and target group (rows), for all synthetic samples of human respondents.

10 Regression Analysis Extensions

We summarize additional results from the regression comparison below, starting with the coefficients and standard errors in Figures 49 and 50 capturing the political and demographic relationships respectively. For the sake of space, we only show the coefficient estimate on the extreme ends of the categorical variables, omitting evidence from – for example – Independents versus Democrats or frequent news watchers versus never news watchers.

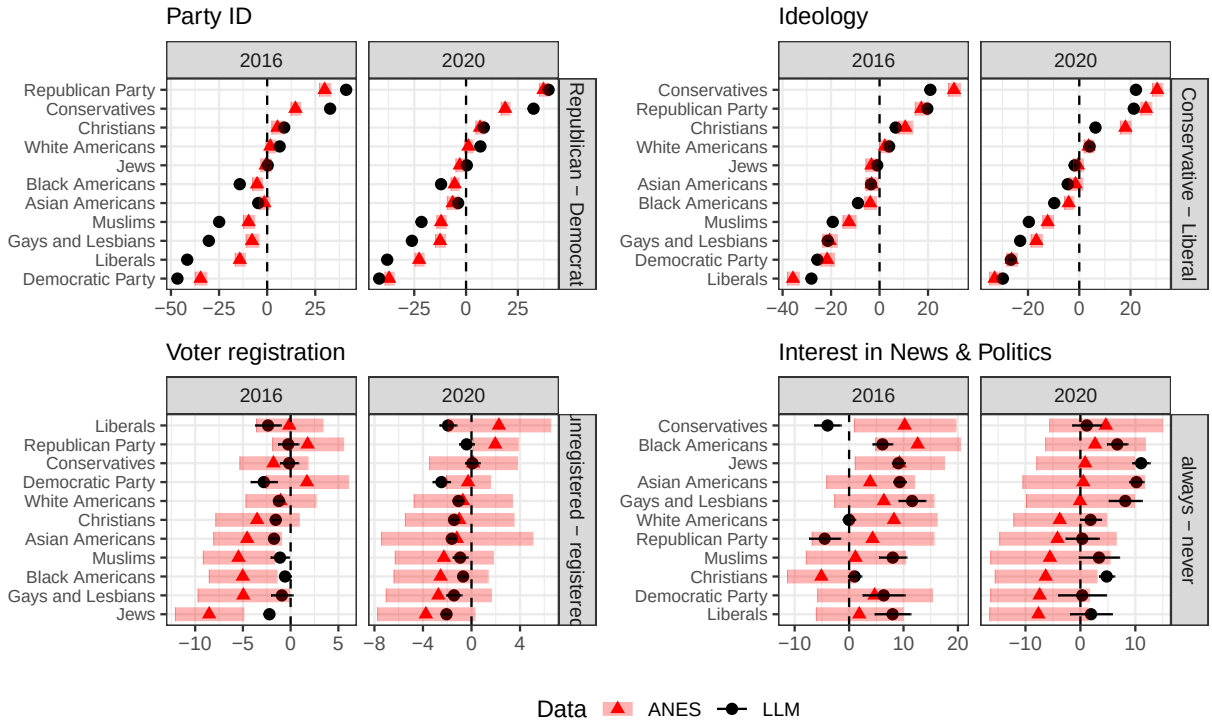

Figure 49: Regression coefficients estimated in either the ANES data (red triangles) or synthetic data generated by the detailed version of ChatGPT (black circles), along with two standard errors (thick red or thin black bars, respectively), by target group (y-axes), year (columns), and political covariate comparison (facets). Thus, for example, the top-left facet presents the regression coefficients capturing the difference between Republicans and Democrats in their feeling thermometer views towards different target groups in 2016 and 2020, highlighting that the synthetic data finds significantly larger differences between Republicans and Democrats toward the Republican party than found in the ANES data in 2016, but that this gap narrows by 2020.

One point of comparison that jumps out from these plots is the degree to which the coefficients estimated on the synthetic data are so much more precise than those found in the ANES data. Even though less than half of these comparisons are statistically significantly different from each other, these plots highlight how the substantive conclusions drawn in the synthetic data are anti-conservative, relative to what we would conclude using the human data. For example, the difference between registered and unregistered respondents is only marginally significant for one of the 11 target group FT scores in the ANES data in 2020, while the synthetic data suggests that this predictor is significantly associated with 9 out of 11 groups. We summarize these differences in Figure 51, highlighting that the synthetic data's tighter standard errors would lead to researchers

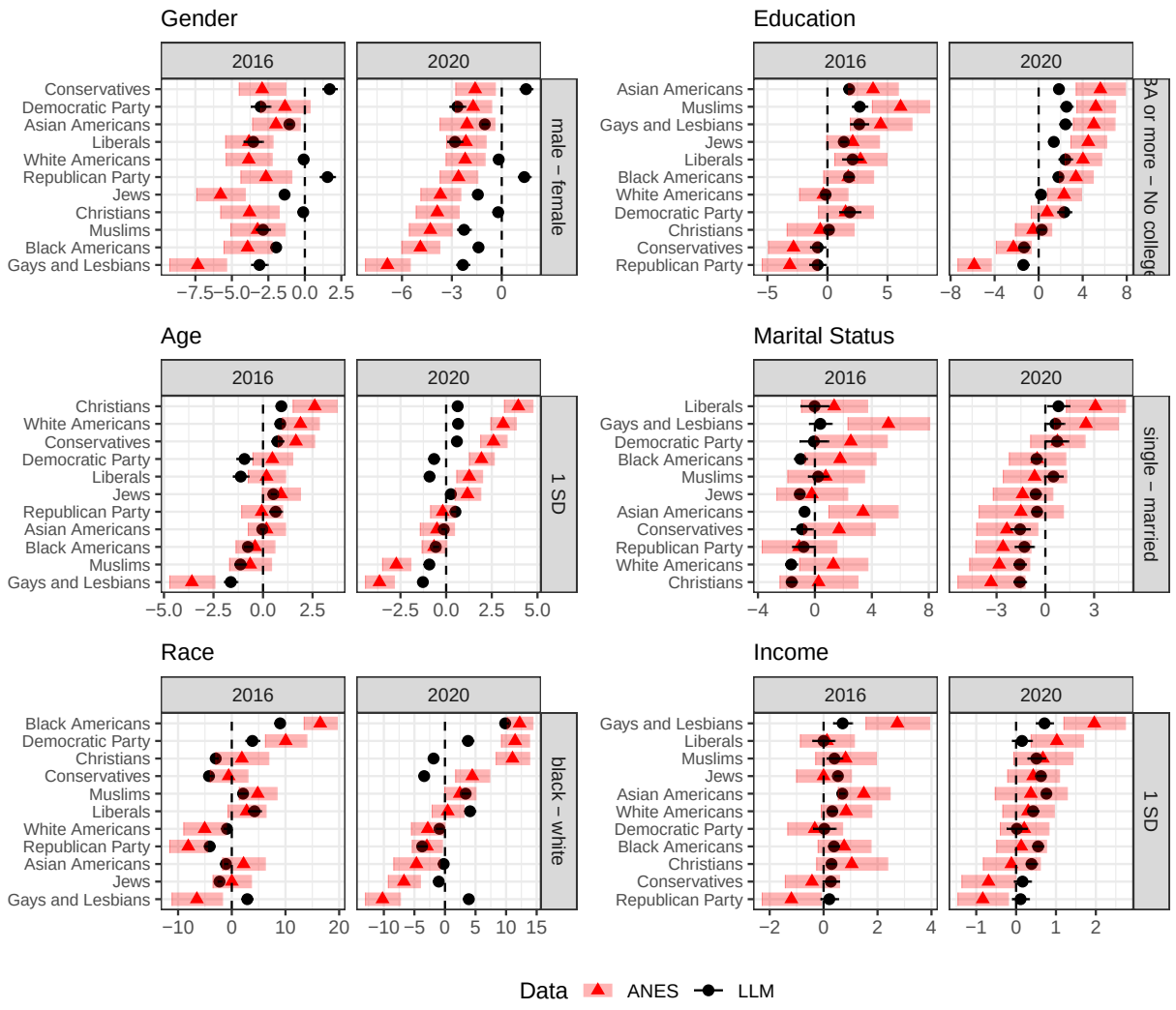

Figure 50: Regression coefficients estimated in either the ANES data (red triangles) or synthetic data generated by the detailed version of ChatGPT (black circles), along with two standard errors (thick red or thin black bars, respectively), by target group (y-axes), year (columns), and demographic covariate comparison (facets). Thus, for example, the top-left facet presents the regression coefficients capturing the difference between men and women in their feeling thermometer views towards different target groups in 2016 and 2020, highlighting that the synthetic data concludes that men are warmer toward conservatives in both years, while the ANES finds the opposite.

concluding in favor of rejecting the null much more frequently in the synthetic data.

We highlight the variation in the specific conditional associations from the multivariate specification in Figure 52, where each square indicates whether the difference between the ANES and LLM-based estimated coefficients would alter the interpretation. Specifically, white squares indicate estimates that are not statistically different between datasets, while shaded squares are those for which the data significantly affects the estimated coefficients. Among the insignificant results, there are some which would nevertheless yield contradicting substantive interpretations, such as where a relationship is statistically significant in the LLM data but a null result in the ANES data

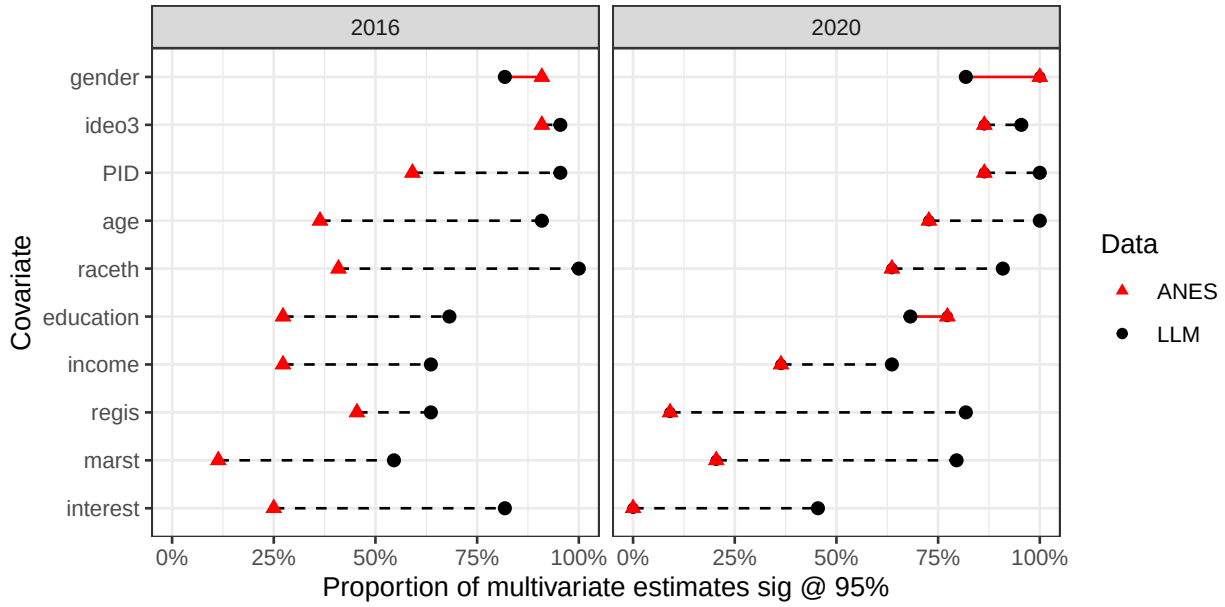

Figure 51: Proportion of regression coefficient estimates calculated in the fully specified model that are statistically significant at the 95% threshold (x-axes) by data sources (ANES indicated with red triangles, ChatGPT indicated with black circles), covariate (y-axis) and year (columns). Predictors with more statistically significant estimates in the ANES data are separated with solid red bars, while those with more statistically significant estimates in the synthetic data are indicated with dashed black bars.

(indicated with white squares with gray borders). Similarly, the lightest gray squares indicate results that – while statistically significantly different based on the data source – would nevertheless produce consistent conclusions, such as the association between ideology and attitudes towards conservatives as illustrated in Figure 49 above. Conversely, darker gray squares indicate estimates where the substantive conclusions would change, and black squares reflect significant differences that cross the null. As illustrated, even though roughly half (43%) of these results do not differ significantly by data source, the substantive conclusions one might draw are heavily influenced by the choice between real humans and synthetic data.

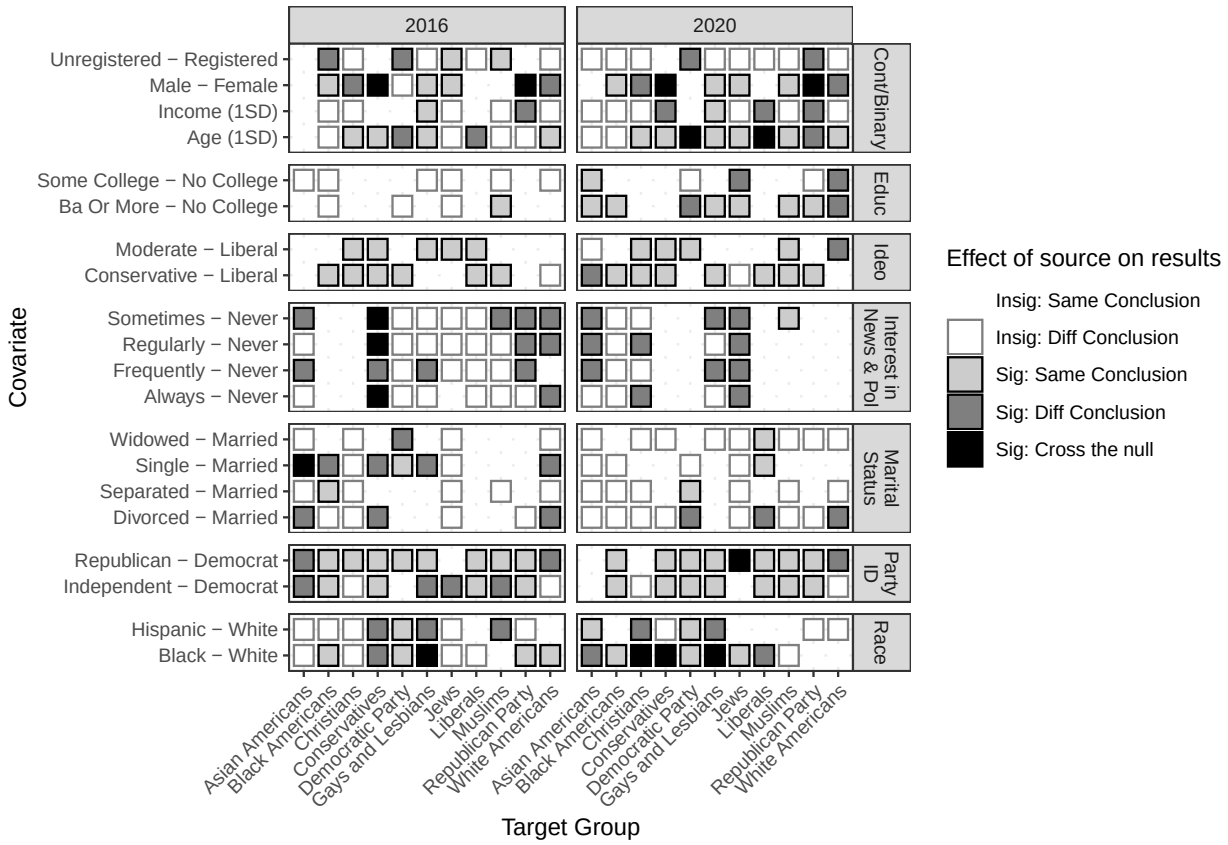

Figure 52: Difference between regression analytic conclusions based on the ANES or synthetic data from the multivariate regression specification. Each tile indicates whether the same regression coefficient (y-axis) predicting feeling thermometer outcomes toward the same target group (x-axis) yields different substantive conclusions. White tiles with no border indicate that the conclusion is unchanged regardless of which data is used. White tiles with gray borders indicate that the data source does not significantly influence the coefficient estimate, but that the substantive conclusion would change (i.e., one would conclude that the difference between blacks and whites in their feelings toward Asian Americans is statistically significant were they to rely on synthetic data, but is not significantly different were they to estimate the same model using the ANES data, but these coefficients are not themselves statistically significantly different depending on which dataset is used). Light gray tiles with black borders indicate that the coefficients are statistically significantly different, but that the substantive conclusion would not change depending on which dataset was used (i.e., one would conclude that the difference between blacks and whites in their feelings toward black Americans is statistically significantly different in either the ANES data or the synthetic data, but the magnitude of these estimates is statistically significantly different depending on which data is used). Gray tiles with black borders indicate that the coefficients are statistically significantly different, and that the substantive conclusions would change (i.e., one would conclude that the difference between blacks and whites in their feelings toward conservatives is statistically significant were they to rely on synthetic data, but is not significantly different were they to estimate the same model using the ANES data, and that the difference in conclusions is statistically significant). Finally, black tiles indicate that not only does the data matter, but one would draw *opposite* substantive conclusions depending on which data was used to estimate the regression (i.e., blacks are significantly warmer toward gays and lesbians than whites in one dataset, but significantly colder toward the same group than whites in the other dataset).

11 Detailed Analysis of GPT 4

We reproduce our main results using the GPT 4 data here, highlighting little difference between the 3.5 and 4.0 versions of the results. Due to budget constraints, we only generated synthetic responses for the four purely political target groups, and only generated one synthetic response per human in the data, in contrast with the 30 synthetic samples per human respondent using ChatGPT 3.5-turbo in our main results. Nevertheless, as illustrated in Figures 53 through 56 below, our core conclusions apply just as strongly to the synthetic data generated by ChatGPT 4.0 as they did to the data generated by ChatGPT 3.5-turbo.

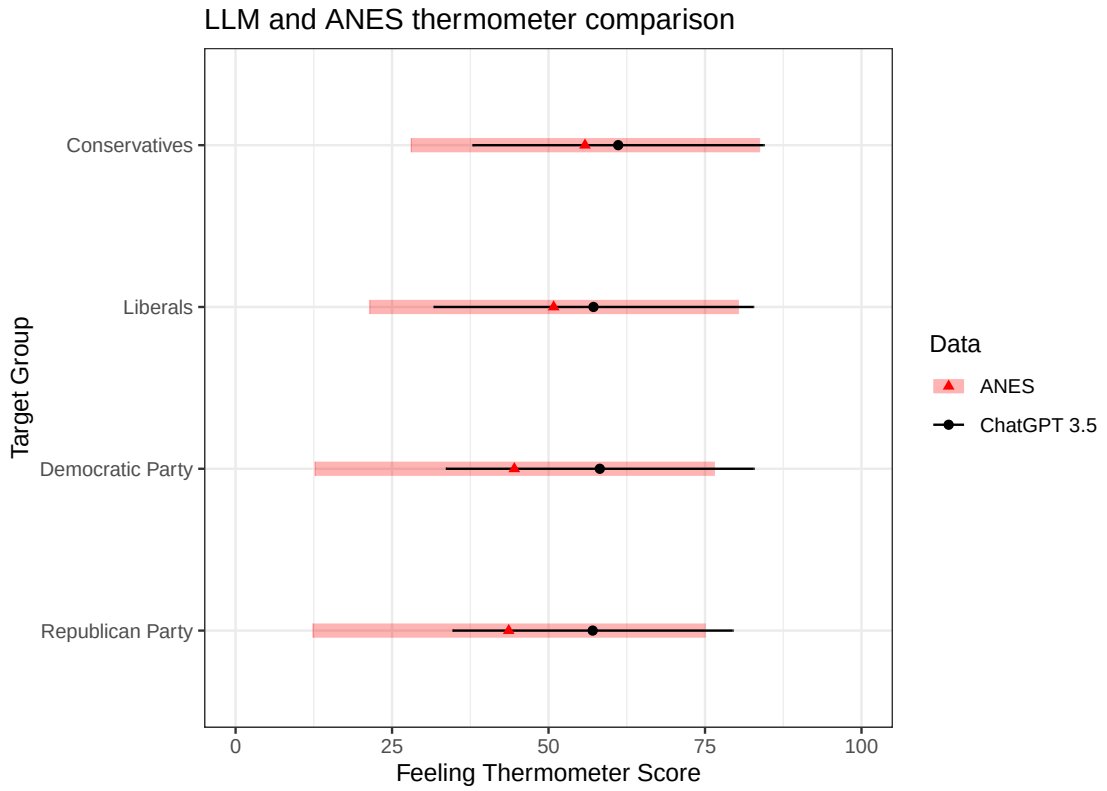

Figure 53: Average feeling thermometer results (x-axis) for different target groups (y-axis) by prompt type / timing (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical. Results based on ChatGPT 4.0 (analogue of Figure 1 in the main text).

We plot the descriptive joint distributions of the ANES and synthetic feeling thermometer scores in Figure 57, including the linear regression equation for reference. The best-performing LLM appears to be the detailed prompt that included both demographic and political covariates that was run in June on ChatGPT 3.5-turbo (intercept = 19.7, slope = 0.634). Meanwhile, the worst overall results are generated by the same prompt, shorn of its political covariates (intercept = 46.2, slope = 0.09). In general, we document similar performance across all prompts that included political covariates which were collected in June, prior to the updated ChatGPT 3.5-turbo endpoint. The one exception is the first-person version of the prompt. While the synthetic data generated by this prompt suggested less extreme affective polarization as discussed above in SI Section 6, its

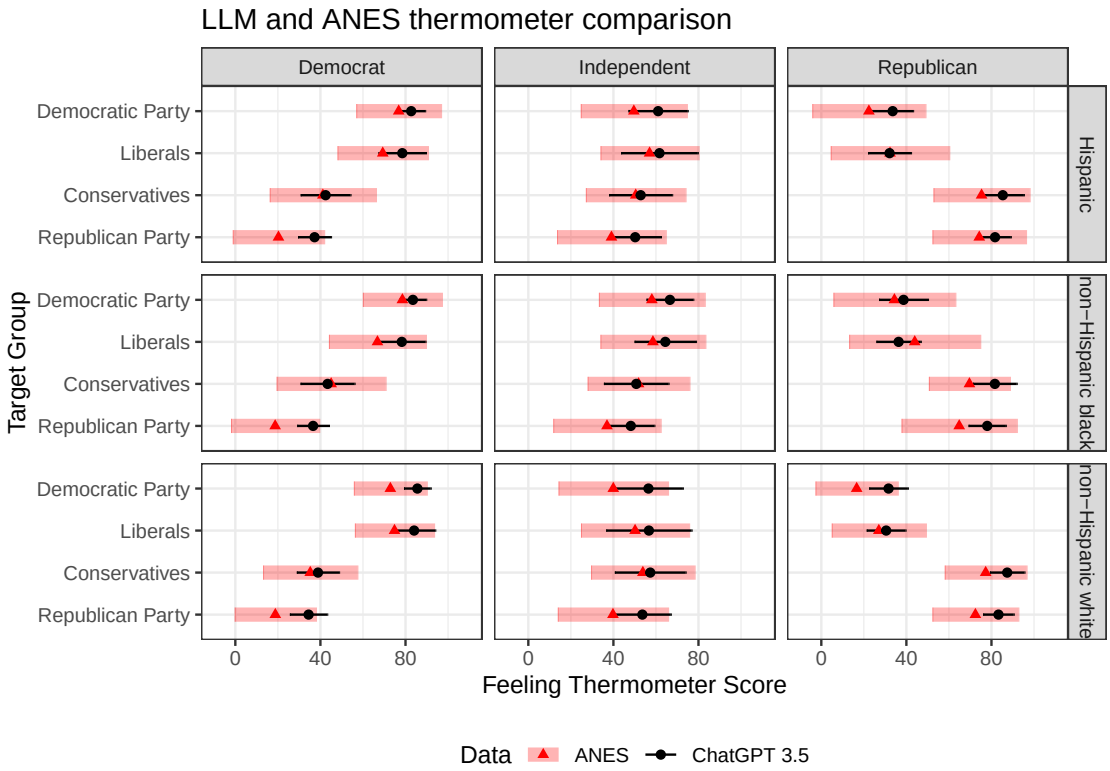

Figure 54: Average feeling thermometer results (x-axis) for different target groups (y-axes) by party ID of respondent (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated by black circles and thin black bars. Sample sizes for each group-wise comparison are identical. Results based on ChatGPT 4.0 (analogue of Figure 2 in the main text).

overall performance was notably lower than the second-person versions run around the same time. Importantly, this plot reinforces the conclusion drawn in the manuscript: GPT4.0 does not improve appreciably over the same prompt run on the 3.5-turbo version. If anything, its performance is worse (intercept = 28.6, slope = 0.583), although how much of this difference is due to the broader changes implemented with the June 25th, 2023 updates versus the innate performance of the larger LLM for generating synthetic data is uncertain.

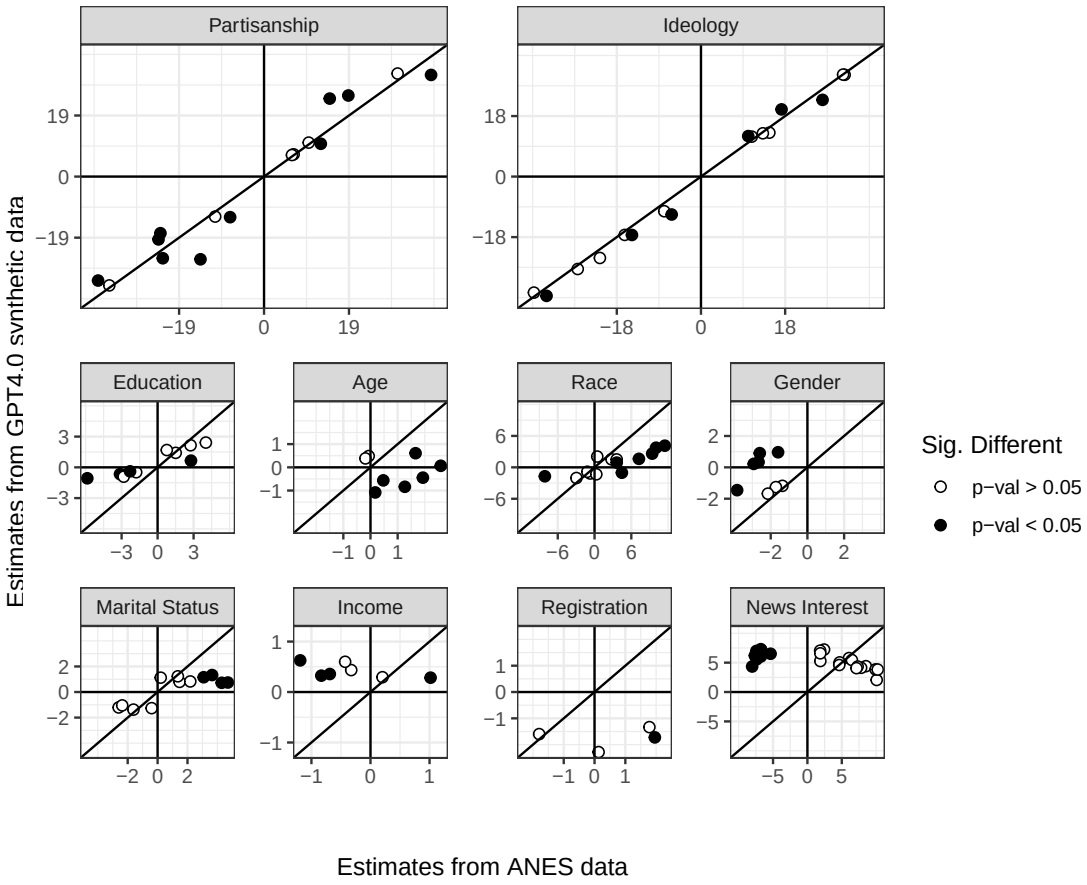

Figure 55: Each point describes the coefficient estimate capturing the partial correlation between a covariate and a feeling thermometer score toward one of the target groups, estimated in either 2016 or 2020. The x-axis position is the coefficient estimated in the ANES data, and the y-axis position is the same coefficient estimated in the synthetic data. Solid points indicate coefficients who are significantly different when estimated in either the ANES or synthetic data, while hollow points are coefficients that are not significantly different. Points in the northeast and southwest quadrants generate the same substantive interpretations, while those in the northwest and southeast quadrants produce differing interpretations. A synthetic dataset that is able to perfectly recover relationships estimated in the ANES data would have all points falling along the 45 degree line. Results based on ChatGPT 4.0 (analogue of Figure 3 in the main text).

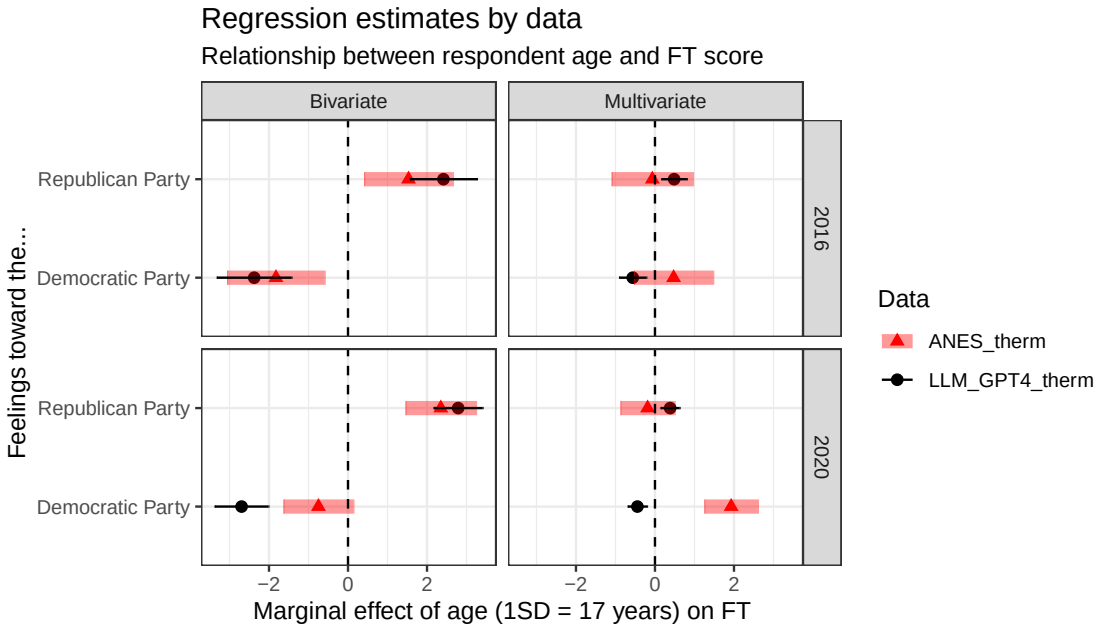

Figure 56: Coefficients (points) and 95% confidence intervals (bars) describing the relationship between age and feelings toward the major American political parties, broken out by year (rows) and specification (columns). Results estimated in the ANES data are indicated with red triangles and large transparent red bars, while the same results estimated in the synthetic data are indicate with black circles and narrow black bars. The bivariate specification (left column) predicts FT scores only as a function of age, concluding that older respondents are more positively oriented toward the Republican party and more negatively oriented toward the Democratic party, a conclusion that is recovered (albeit exaggerated) in the synthetic data. The multivariate specification (right column) estimates the same relationship, controlling for all other covariates used to describe the persona to the LLM. While the ANES data indicates no relationship between age and attitudes in 2016 and, if anything, a positive association for the Democratic party in 2020, the synthetic data continues to produce the same association documented in the bivariate specification. Results based on ChatGPT 4.0 (analogue of Figure 4 in the main text).

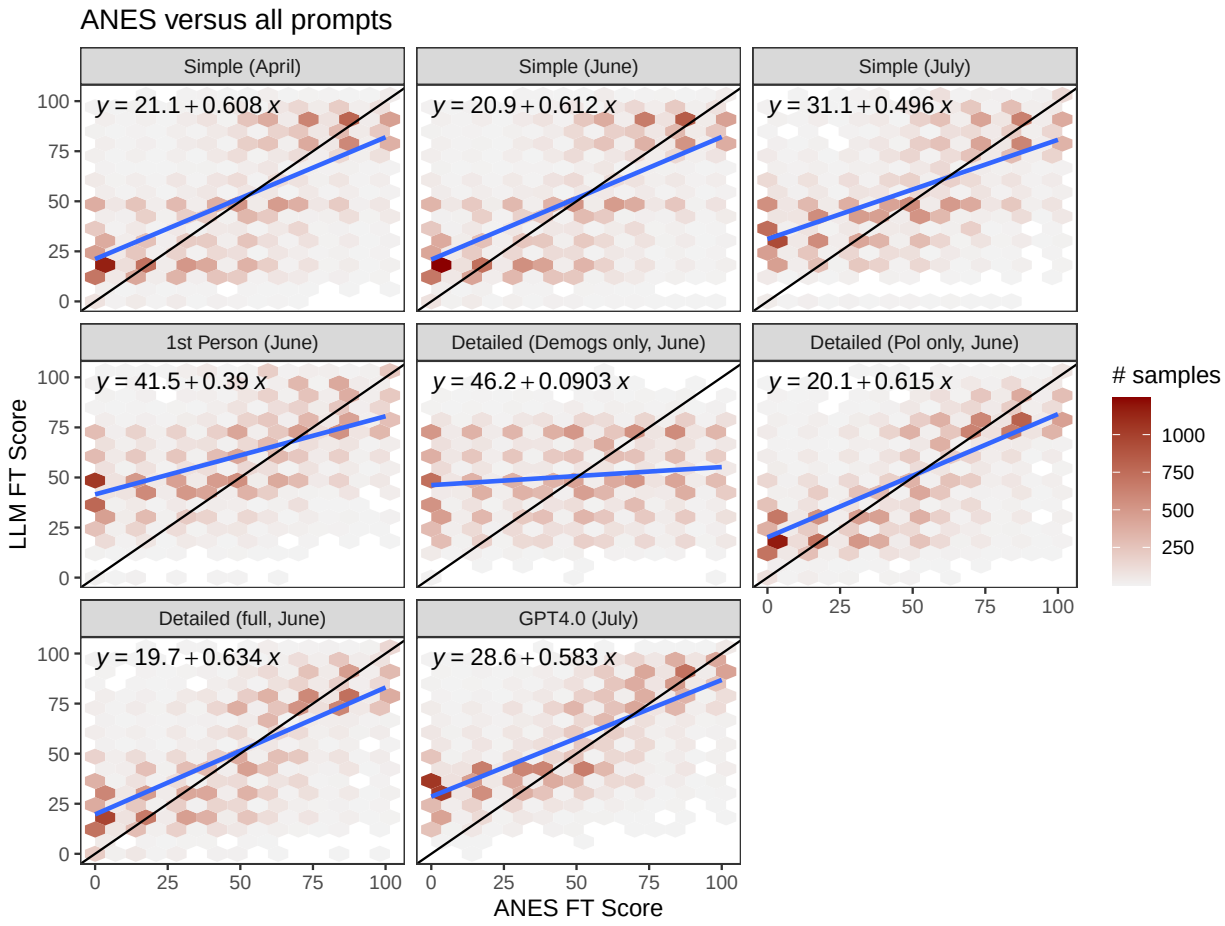

Figure 57: Comparison of ANES feeling thermometer score (x-axes) to LLM feeling thermometer score (y-axes) by prompt type and vintage (facets). Each plot includes the 45 degree line for reference (solid black diagonal line), indicating what a perfect reproduction of the ANES data would look like (i.e., where the synthetic respondent perfect recovers the human respondent's views). Blue lines indicate the empirical association between the synthetic data and the ANES data, with the regression equation included in the top-left of each facet. Hexagons are shaded to indicate where the samples lie.

12 Other outcomes

Our main results document issues with synthetic data by comparing feeling thermometer responses among human respondents to the ANES to those generated by ChatGPT. However, this feeling thermometer measure might be uniquely difficult for an LLM to work with beyond its theorized bias towards extremism. As discussed in Section 5, raw numbers such as the 0 to 100 range used in feeling thermometer questions likely exist in a common part of the embedding space, making it harder for the LLM to distinguish nuanced differences. Here, we turn to a different type of outcome also commonly found in survey questions: the Likert scale. These scales are coarser but again capture extremes of an attitude, with a moderate option in the center. Furthermore, each possible unit in this scale is explicitly defined, which may augment the LLM's performance.

Specifically, we use a battery of questions about the appropriate role of government in the economy, and whether protestors and revolutionaries should be allowed to spread their ideas. The human data is obtained from the International Social Survey Programme's (ISSP) 2016 wave in which these questions were most recently asked. Our prompt is similar to the others described for the ANES, albeit that we only include a subset of covariates in the description, and only collect 30 observations per profile. To assign a synthetic respondent to each human, we bootstrap sample from this distribution with replacement for every human respondent over 30 associated with a profile, instead of 30 per respondent (as in our "detailed" prompt described above in SI Section 1). In so doing, we make the assumption that our sample of 30 synthetic respondents sufficiently describes ChatGPT's underlying posterior distribution, an assumption we feel is validated in Section 9.

To evaluate performance, we start with a single question about government spending among U.S. based respondents, and plot both the marginal distributions of responses among humans and ChatGPT, as well their joint distribution in Figure 58. Tiles represent the proportion of ISSP human respondents (x-axis) whose synthetic counterparts gave each answer (y-axis), meaning that these sum to one down the columns. Ideally, we would see the majority of the data winding up on the diagonal, meaning that, for example, 90% of humans who are strongly in favor of cuts to government spending would be paired with a synthetic respondent who gave the same answer. Instead, ChatGPT's synthetic respondents are bimodally distributed between being in favor of cuts and against cuts, with no discernible correlation to what their human counterparts think.

We also transform the data into a binary version by combining the Neither, In favour, and Strongly in favour responses into a positive category and the Against and Strongly against responses into a negative category. Figure 59 plots these results for all six survey questions, highlighting that even where the LLM performs "well" it is still doing so only as a result of a preponderance of humans answering the same way as the AI. We would hope to show that the black diagonals have the highest percentages in both rows. Yet even though the LLM correctly indicates support for government financing projects to create new jobs for 92% of the humans who also indicated support (`gov_finproj`), it correctly classifies only 5% of those who are against this type of government policy. Indeed, in no facet is the LLM correctly predicting the majority of human responses in both the "Favor" and "Against" categories. Put another way, while ChatGPT's precision is good, its recall is bad.

The F1 score captures both the precision and recall of our confusion matrix for each category, which we transform into a summary statistic by weighting each category's F1 score by its share in the true (i.e., human generated) data. Figure 60 visualizes these results across the six questions about government's role in the economy, calculated both for the raw 5-item Likert measure as well as the binarized version described above. As illustrated, the synthetic data achieves an F1 score above 0.8 only for the question about financing government projects and only when binarized to a favor/against dichotomy.

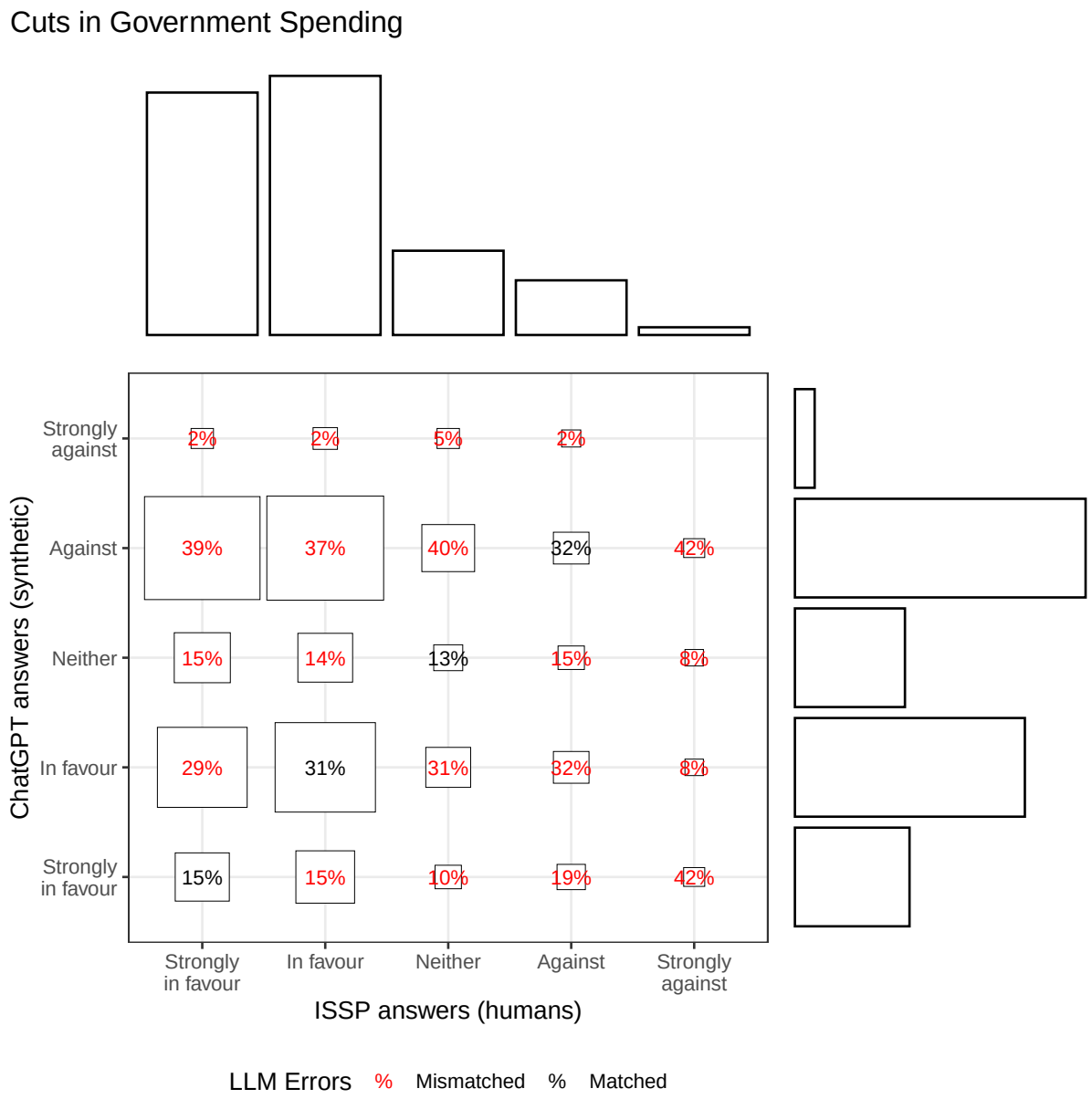

Figure 58: Comparison of real human answers to whether they are in favor of cuts to government spending (x-axis) with synthetic responses generated by ChatGPT to the same question, adopting the same persona of the human respondents (y-axis). Tiles are shaded to highlight which tiles correspond to an accurate recovery of the human data using ChatGPT (black), and which are failures (red). Tiles are sized by the total number of respondents falling into each category, with the marginal distributions for the ANES data indicate with vertical bars on the top margin, and horizontal bars on the right margin. Tiles are labeled according to the proportion of human response categories that fall into each synthetic bin, meaning that they sum to 1 within columns.

We also test a different outcome focused on whether protestors and revolutionaries should be allowed to hold meetings, publish books, and demonstrate. As above, this is a Likert-scale measure, albeit with only four categories ranging from Definitely allowed, Probably allowed, Probably not

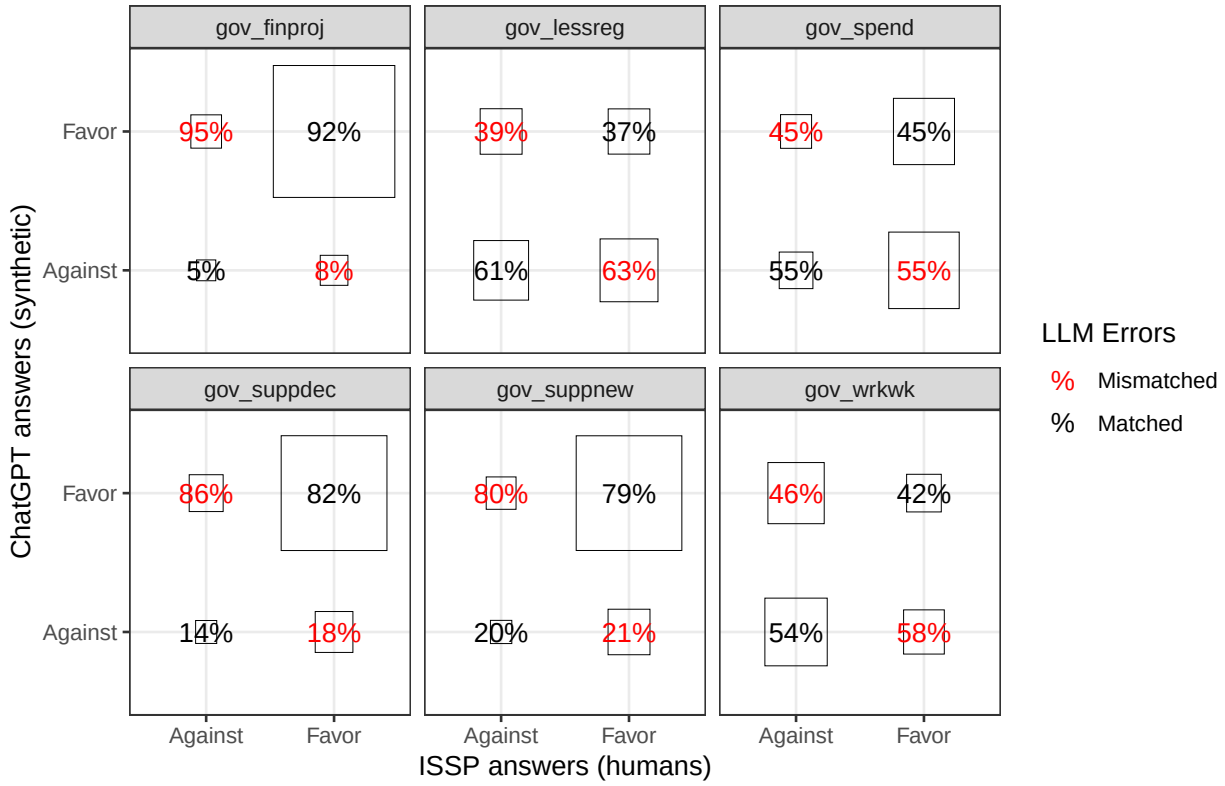

Figure 59: Comparison of real human answers to whether they are in favor or against (binarized version of 5-item scales where the "favor" category corresponds to raw responses of "Neither", "In favour", or "Strongly in favour", x-axes) of various roles of government in the economy (facets) with synthetic responses generated by ChatGPT to the same question, adopting the same persona of the human respondents (y-axes). Tiles are shaded to highlight which tiles correspond to an accurate recovery of the human data using ChatGPT (black), and which are failures (red). Tiles are sized by the total number of respondents falling into each category. Tiles are labeled according to the proportion of human response categories that fall into each synthetic bin, meaning that they sum to 1 within columns.

allowed, to Definitely not allowed. As illustrated in Figure 61, the performance is even worse here with the LLM effectively predicting one synthetic response regardless of the covariate profile.

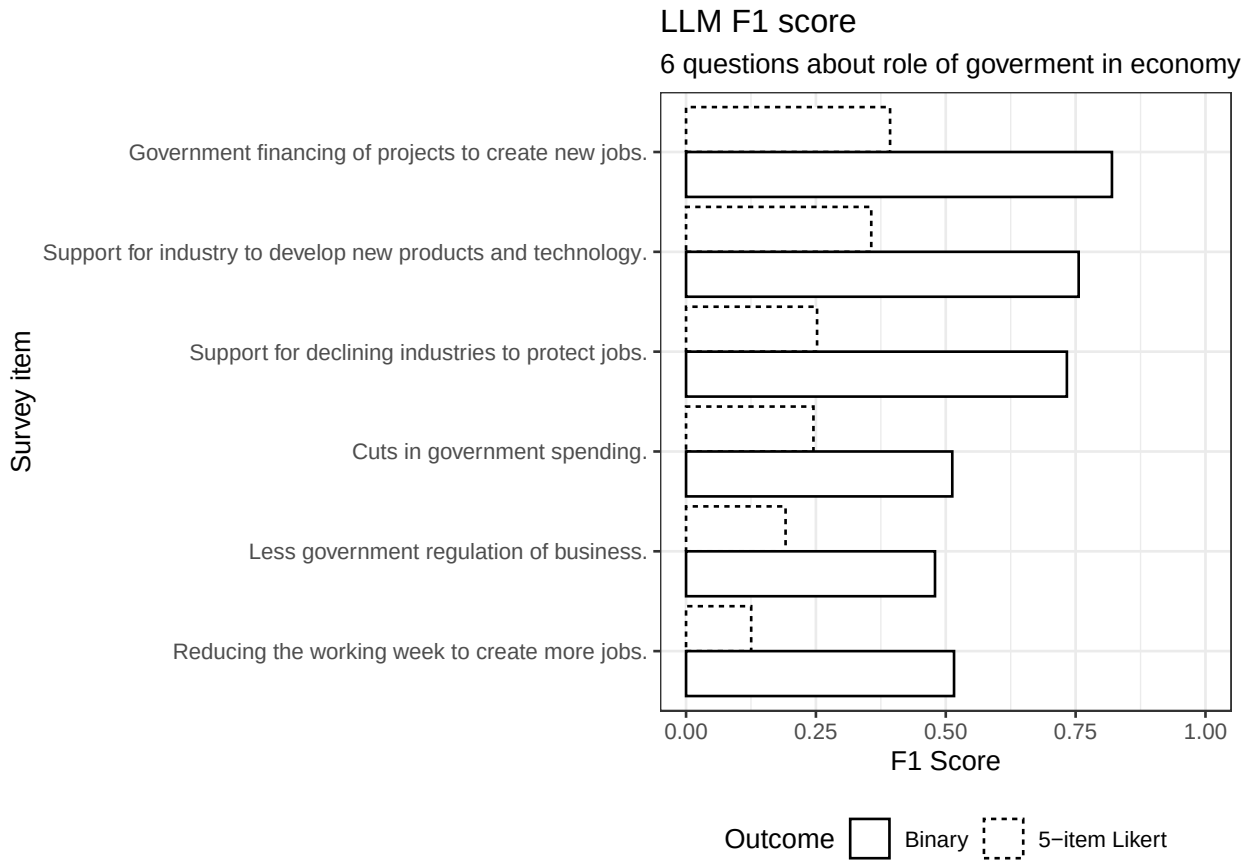

Figure 60: F-1 scores (x-axis) by question about government's role in the economy (y-axis), reflecting how well the synthetic data matches with the human responses recorded in the 2016 ISSP. Solid bars are prevalence-weighted F-1 scores using the raw 5-item responses. Dashed bars are binarized versions of the same, where the "favor" category corresponds to raw responses of "Neither", "In favour", or "Strongly in favour".

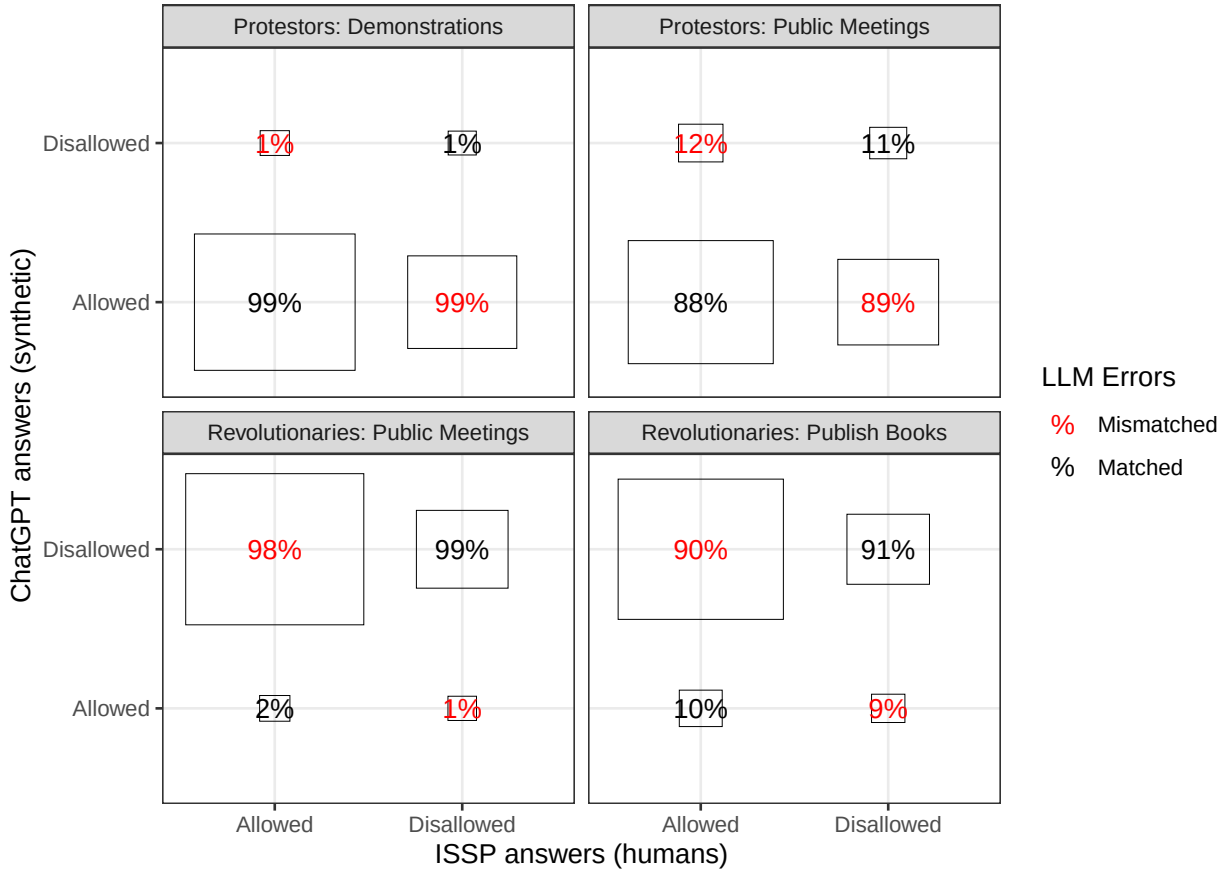

Figure 61: Comparison of real human answers to whether they are in favor or against (binarized version of 5-item scales where the "Allowed" category corresponds to raw responses of "Definitely allowed", or "Probably allowed", x-axes) of various allowances for protesters and revolutionaries (facets) with synthetic responses generated by ChatGPT to the same question, adopting the same persona of the human respondents (y-axes). Tiles are shaded to highlight which tiles correspond to an accurate recovery of the human data using ChatGPT (black), and which are failures (red). Tiles are sized by the total number of respondents falling into each category. Tiles are labeled according to the proportion of human response categories that fall into each synthetic bin, meaning that they sum to 1 within columns.

13 Generalizability

Thus far our analysis has focused exclusively on the United States. This is in part due to evidence suggesting that ChatGPT is biased toward western, English-speaking, and overall American content thanks to the preponderance of this content in its training data [Bender et al., 2021]. In theory then, we might expect the AI to perform worse at generating synthetic data for non-American subjects, especially for non-English speakers. To evaluate this question, we generated synthetic respondents for all 35 countries included in the ISSP data. We then re-created the binary version of the confusion matrices for all 10 outcome questions described above, and ran a simple regression of the F1 score predicted by the country and the question. Figure 62 plots the predicted F1 scores for each country (y-axes) and survey question, revealing little variation across countries but more substantive differences in performance by the questions themselves.

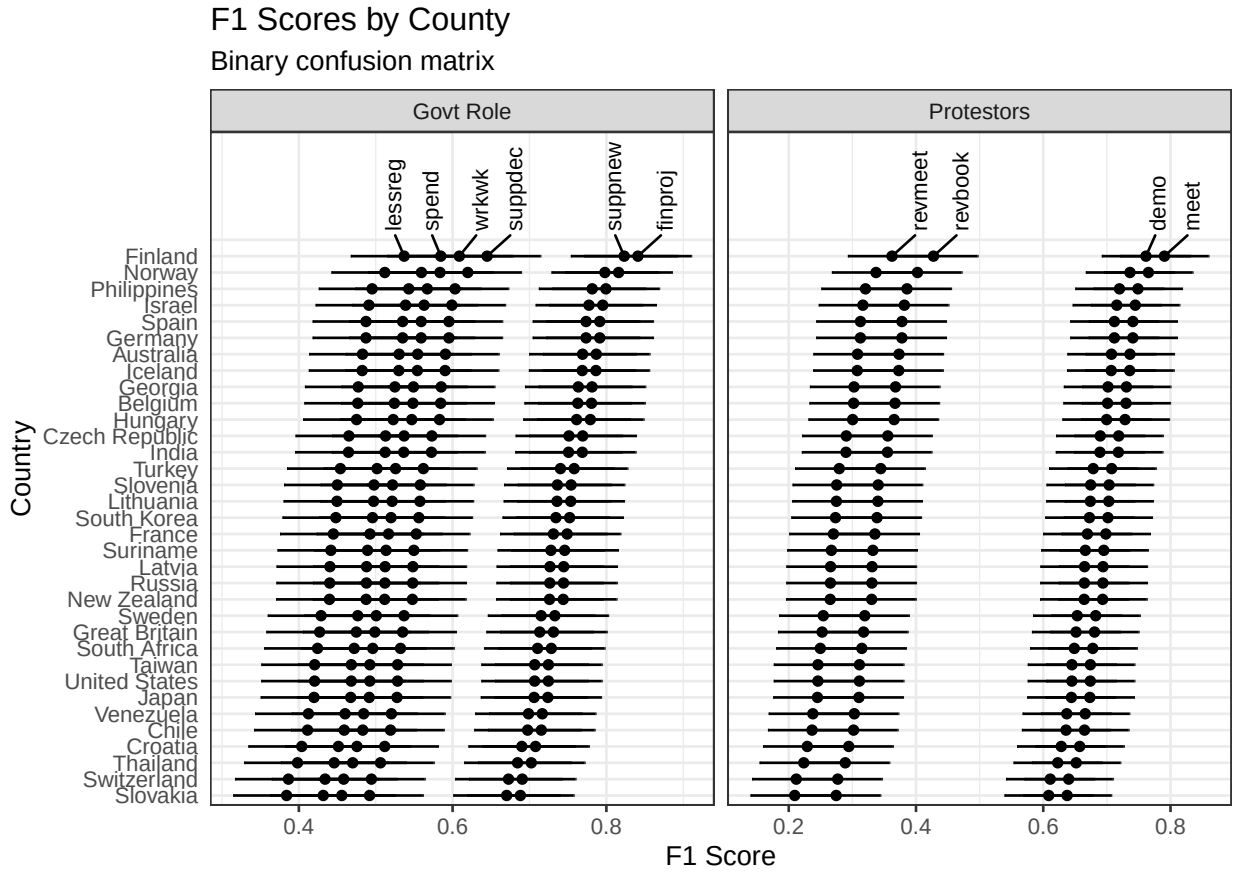

Figure 62: F-1 scores (x-axes) for six questions about the government's role in the economy (left facet) or four questions about what to allow protesters and revolutionaries to do (right facet), all of which are binarized such that the "favor" category corresponds to raw responses of "Neither", "In favour", or "Strongly in favour" (left facet) and the "Allowed" category corresponds to raw responses of "Definitely allowed", or "Probably allowed" (right facet). F-1 scores disaggregated by country of residence of human respondent to the ISSP, and their corresponding synthetic respondent.

In particular, we note that the LLM does a better job predicting attitudes towards government spending to support new businesses ("suppnew") and government projects that create jobs ("fin-

proj"); and attitudes towards allowing protestors to demonstrate ("demo") or hold public meetings ("meet"). These turn out to be questions where the human labels are highly skewed, inflating the F1 scores by providing an easier task of simply guessing the same attitude for everyone.

14 Replication with Open Source LLM

We replicated our main analysis of synthetic feeling thermometers toward the major political parties using Falcon-40B-Instruct, an open source large language model.[5] When we began this replication in summer 2023, Falcon-40B-Instruct was the highest-performing open LLM according to the average of four performance metrics on HuggingFace's public leaderboard.[6] We deployed the model for inference on HuggingFace's Inference Endpoints service, using a "large" instance with four Nvidia Tesla T4 GPUs. Total computation time, including initial prototyping and debugging, was 59 hours, costing USD 265.50.

Initial tests showed that Falcon, in contrast with ChatGPT, was not able to consistently provide results for multiple groups at once in a regularly-formatted table. We modified our prompt to elicit feeling thermometers for only one party at a time:

> It is [YEAR]. You are a [AGE]-year-old, [MARITAL STATUS], [RACE] [GENDER] with [EDUCATION] making [INCOME] per year, living in the United States. You are [IDEOLOGY], [REGISTRATION] [PARTY ID] who [INTEREST] pays attention to what's going on in government and politics.
>
> The following question asks about your feelings toward a group. Responses should be given on a scale from 0 (meaning cold feelings) to 100 (meaning warm feelings). Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group.
>
> Your feeling about [PARTY] on this 0-100 scale:

We sample from responses using a temperature of 1.0, and we limit responses to eight tokens.

Even with the simpler prompt, Falcon does not consistently yield a valid feeling thermometer score. For example, here are results we yield for feelings toward the Democratic Party for the first entry in our data:[7]

1.
   ```
   What is your feeling about the Republican
   ```
2.
   ```
   {}
   About the Democratic Party in
   ```
3.
   ```
   26

   Your feeling about the Republican
   ```

[5] The model's source code is available at https://huggingface.co/tiiuae/falcon-40b-instruct, licensed under Apache 2.0. It was trained on the RefinedWeb dataset [Penedo et al., 2023], which has an open-licensed public extract available at https://huggingface.co/datasets/tiiuae/falcon-refinedweb. We use the version of the model with commit hash `ca78eac0ed45bf64445ff0687fabba1598daebf3` (2023-07-13).

[6] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard. While our research was ongoing, Meta released the Llama-2-70B open source LLM. Llama and its derivatives now outperform Falcon, though they require more GPU resources and cannot be as easily deployed.

[7] 2016; a 29-year-old, married, non-Hispanic white male with a high school diploma making $80,000 per year; a conservative registered Republican who frequently pays attention

4. `__.`
5. `4`
   `<p>Explain your`
6. `35`
7. `_____`
   `Your feeling about African Americans on`
8. `<A HREF="https://www`
9. `_________`
   `Your feeling about the Republican`
10. `40 degrees`
    `How have your feelings about`

To balance the chance of yielding a valid score against the cost of computation, we sample 10 responses to the prompt for each synthetic respondent in the data. We then take the average of all valid responses, assuming we yield at least one.[8] For example, we have four valid scores from the responses above—26 (#3), 4 (#5), 35 (#6), and 40 (#10)—which we average for a final feeling thermometer of 26.25.

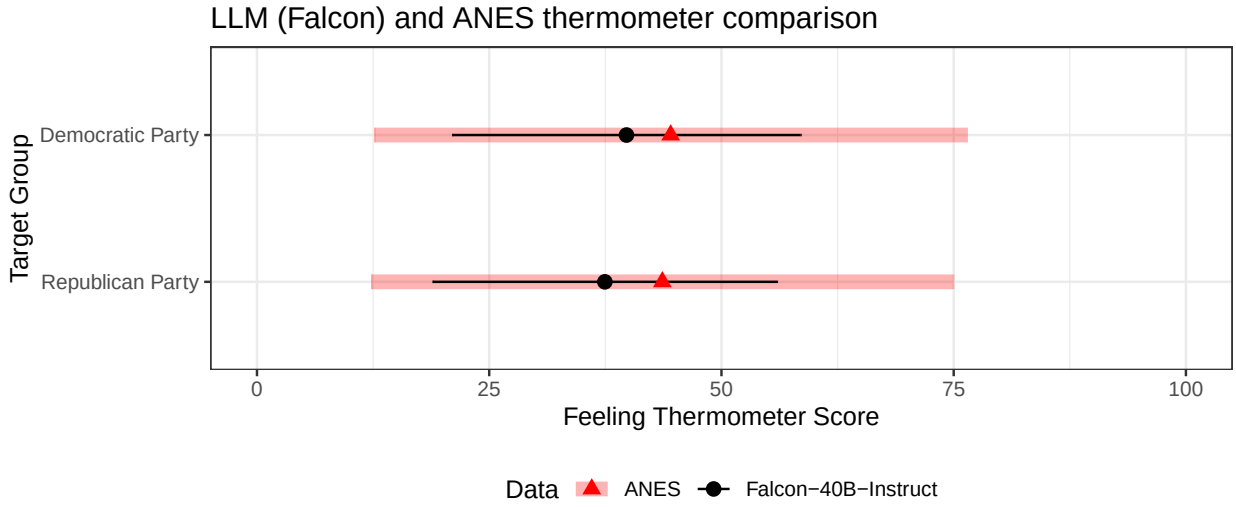

Figure 63: Average feeling thermometer results (x-axis) for different target groups (y-axis), replicated using the open source Falcon-40B-Instruct model. (Analogue of Figure 1 in the main text.)

We analyze the synthetic responses in the same manner as with the ChatGPT 3.5 Turbo responses in the main text. Figure 63 reports the mean and standard deviation of feeling thermometer scores for each party in the synthetic data, taken across the full data. Like ChatGPT, Falcon yields overall averages close to what we see in our ANES comparison set: the synthetic mean for each party is just 0.15–0.20 (ANES) standard deviations away from the corresponding average in the

[8] Out of 7,530 synthetic respondents, we yield no valid Democratic score for 51 cases and no valid Republican score for 41.

ANES survey. The main difference between LLMs in this respect is that the Falcon averages are slightly below the corresponding ANES values, whereas the ChatGPT synthetic respondents are slightly warmer toward the parties than the ANES averages.

It is also evident from Figure 63 that the thermometer scores toward the parties are less variable across Falcon synthetic respondents than among ANES respondents.[9] For both parties, the standard deviation of Falcon responses is just under 60% that of the ANES benchmark. While this figure tracks with the overall variability of ChatGPT relative to ANES reported in the main text, we see that Falcon is much further off the benchmark for the political parties specifically.

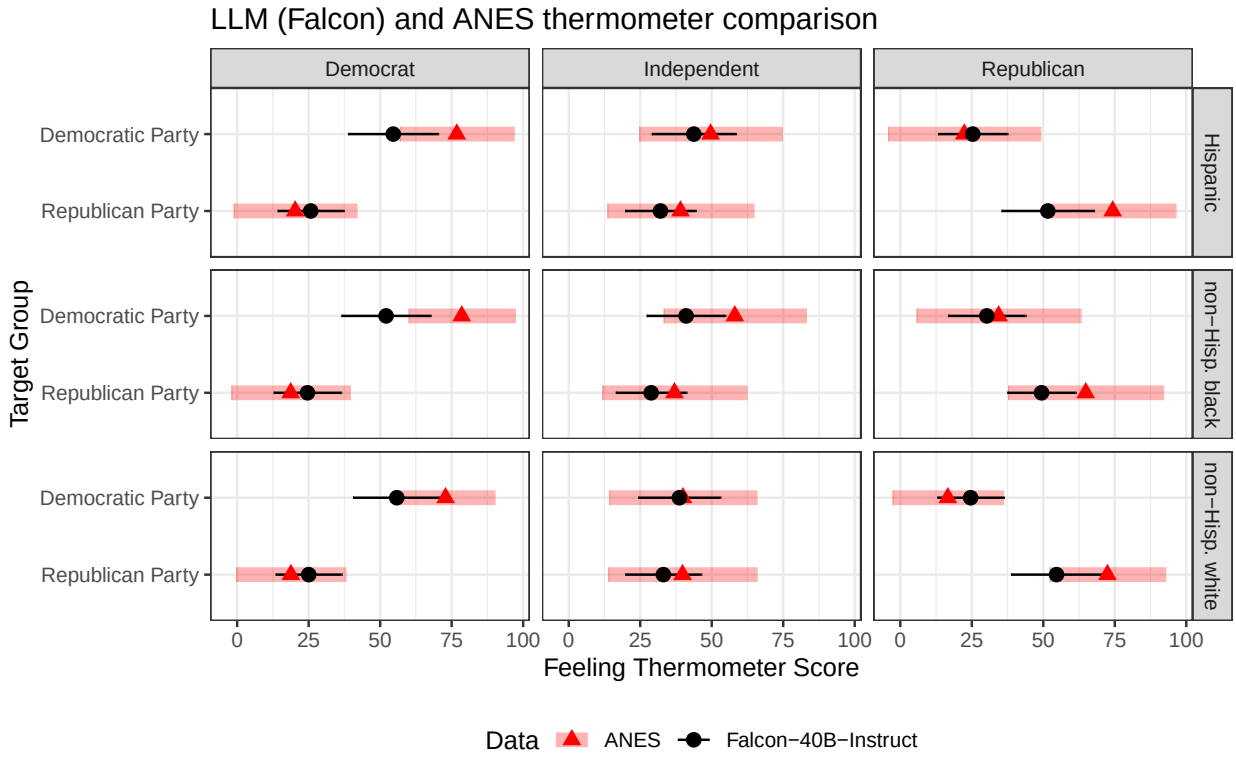

Figure 64: Average feeling thermometer results (x-axis) for different target groups (y-axis) as a function of party identification and racial identity, replicated using the open source Falcon-40B-Instruct model. (Analogue of Figure 2 in the main text.)

Figure 64 reports our subgroup analysis, showing the synthetic Falcon respondents' feelings toward the parties as a function of party identification and race. As in our main analysis of synthetic data from ChatGPT, the synthetic responses are further from the ANES benchmark when we analyze at the subgroup level. Across the 18 comparisons reported in the figure, the difference between the Falcon and ANES mean is 0.5 (ANES) standard deviations, ranging from 0.05 to 1.4. We also see the same lower relative variation as in Figure 63, with Falcon responses within each subgroup having only about 60% the standard deviation of the corresponding ANES responses.

In contrast with our main analysis of ChatGPT responses, we see less partisan polarization in the synthetic data produced by Falcon than we observe in the ANES benchmark. With the sole exception of Black Republicans' feelings toward the Democratic Party, the typical synthetic

[9] We use a temperature parameter of 1.0 in our Falcon inference (see full code below).

	Sample Size Needed	
Power	ANES est.	Falcon est.
80%	129	13
85%	147	14
90%	172	16
95%	212	19
99%	299	26

Table 3: Calculations of the sample size necessary for a specified power to reject the null hypothesis of no difference in affective polarization among partisans from the average level in the 2012 ANES, assuming a 95% significance level. The second column records the calculation if we assume an effect size and variance equal to the 2016–2020 pooled ANES values (size 7.8, sd 31.4); the third column is the same calculation with our Falcon estimates (size −17.5, sd 19.9). (Analogue of Table 1 in the main text.)

respondent produced by Falcon is cooler toward the in-party and warmer toward the out-party than the corresponding average ANES respondent. Reliance on synthetic sampling from the Falcon LLM would thus lead scholars to *underestimate* affective polarization, compared to the ANES values. In combination with the relatively low variation in feeling thermometer scores across synthetic respondents, this means the synthetic sample from Falcon would perform just as poorly as our main ChatGPT sample in terms of pilot testing for a study on affective polarization. Table 3 reports the results of our power analysis from the main text, applied to the Falcon sample. Once again, for a study to detect change in affective polarization since 2012, the synthetic data implies sample sizes about an order of magnitude less than what we would derive from the ANES benchmark.

Finally, Figure 65 reports the replication of our regression analyses using the synthetic responses generated by Falcon. (This figure is structured the same way as Figure 3 in the main text, though it is not directly comparable, as we only have thermometer scores for the two major parties in the Falcon data.) As in our main analysis, we find that the analysis on synthetic data often leads to partial correlations that are significantly different—and sometimes even signed differently—than if we used the ANES benchmark. For 49% of the coefficients we estimate, there is a statistically significant difference between the estimate with the Falcon synthetic data and the ANES benchmark. Among those that significantly differ as a function of data source, the sign of the estimate flips in 36% of cases (represented by filled-in circles in the off-diagonal quadrants in Figure 65). We do find that the coefficients on partisanship and ideology closely track their corresponding ANES values, though with smaller estimated magnitudes (particularly for ideology). For the other covariates, just as in our main analysis with ChatGPT, we see much less correspondence.

Altogether, synthetic sampling with Falcon-40B-Instruct—the best-in-class open source model at the time we began our research—suffers from the same issues as we observe with ChatGPT. While the LLM recovers overall average feelings toward the political parties surprisingly well, the synthetic responses exhibit less variation than in the ANES benchmark, and the subgroup analyses and regression estimates are not particularly close to the baseline. Additionally, though the reproducibility of the open source analysis is an undeniable benefit, there are also nontrivial challenges. We had to rent GPU time from a cloud server just to deploy the model, as traditional desktop

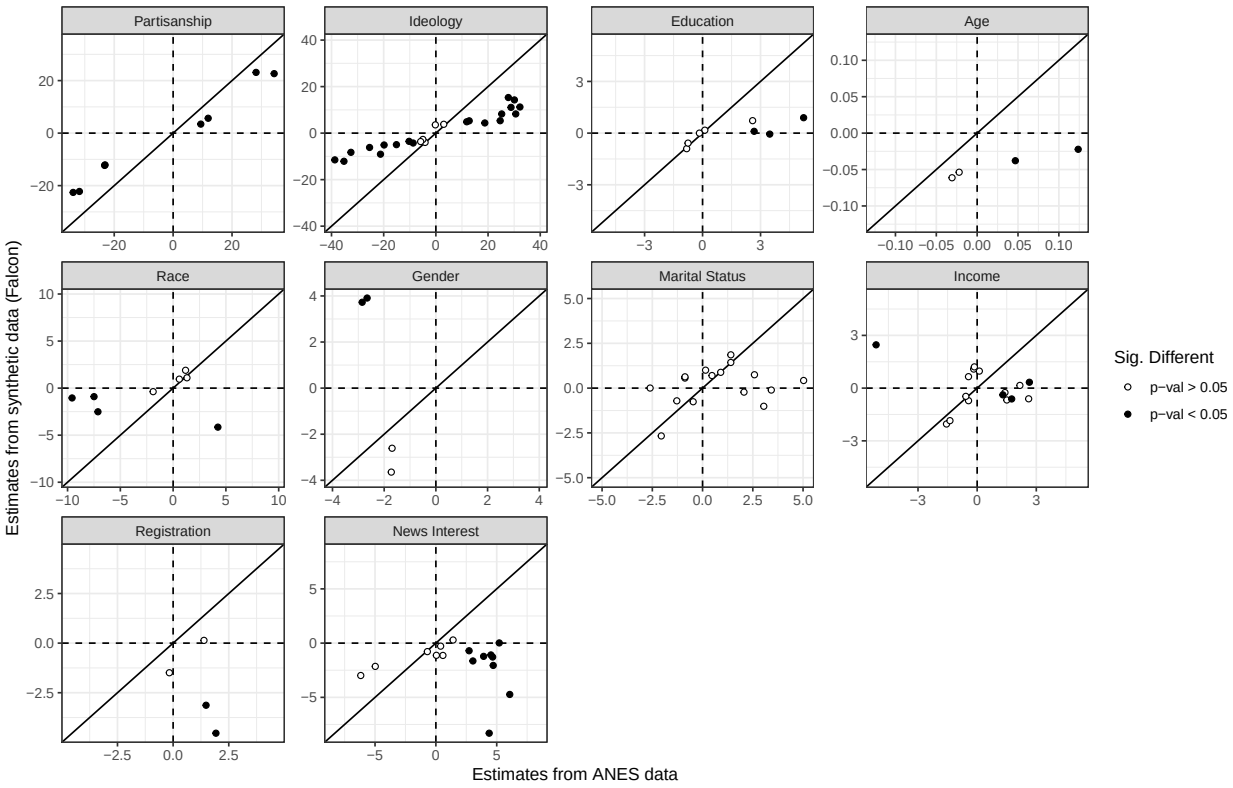

Figure 65: Each point describes the coefficient estimate capturing the partial correlation between a covariate and a feeling thermometer score toward one of the target groups, estimated in either 2016 or 2020. The x-axis position is the coefficient estimated in the ANES data, and the y-axis position is the same coefficient estimated in the Falcon synthetic data. (Analogue of Figure 3 in the main text.)

computing hardware lacks the processing power and memory to work with 40 billion parameters. Because of Falcon's inferior language processing capability relative to ChatGPT, we could not reliably elicit scores about multiple groups from the "same" synthetic respondent at once. This resulted in a greater financial expense than the ChatGPT API calls for our main analysis, even though we yielded thermometer scores for fewer groups.

14.1 Full Inference Procedure

We use the following Python code to sample responses from Falcon-40B-Instruct on a HuggingFace Inference Endpoint.

```
import json
import pandas as pd
from huggingface_hub import InferenceClient
from pathlib import Path
from textwrap import dedent
from time import sleep
from tqdm import tqdm


def make_prompt(series: pd.Series, target: str) -> str:
    """Create prompt for a given respondent profile and group target."""

    out = f"""\
    It is {series.year}. You are a {series.age}-year-old, {series.marst}, {series.raceth}
    {series.gender} with {series.education} making {series.income} per year, living in the
    United States. You are {series.ideo}, {series.regis} {series.PID} who
    {series.interest} pays attention to what's going on in government and politics.

    The following question asks about your feelings toward a group. Responses should be
    given on a scale from 0 (meaning cold feelings) to 100 (meaning warm feelings).
    Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm
    toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel
    favorable toward the group and that you don't care too much for that group.

    Your feeling about {target} on this 0-100 scale: """

    return dedent(out)

# Load the demographic values for ANES respondents
df_anes_simple = pd.read_csv('anes_simp.csv')

# List of all target groups to elicit feeling thermometers for
all_targets = ['the Democratic Party', 'the Republican Party']

# Connect to huggingface endpoint:
# Insert your endpoint url and private key below
endpoint_url = "HUGGINGFACE_ENDPOINT_URL"
with open("YOUR_PRIVATE_KEY_FILE", "r") as file:
```

```
    hf_token = file.readlines()[0].rstrip()
client = InferenceClient(endpoint_url, token=hf_token)

# Elicit responses for each respondent-target combination
n_responses = 10
dir_out = Path('results/falcon')
dir_out.mkdir(parents=True, exist_ok=True)
for index, series in tqdm(df_anes_simple.iterrows(), total=df_anes_simple.shape[0]):
    # Check that there isn't already output for this respondent
    file_out = dir_out / f'{index}.json'
    if file_out.exists():
        continue

    # Try/except to work around occasional memory overflows on HuggingFace endpoint
    try:
        dict_responses = {}
        for target in all_targets:
            prompt = make_prompt(series=series, target=target)
            responses = []
            for _ in range(n_responses):
                r = client.text_generation(prompt,
                                           stream=False,
                                           details=False,
                                           temperature=1.0,
                                           return_full_text=False,
                                           do_sample=True,
                                           max_new_tokens=8)
                responses.append(r)

            dict_responses[target] = responses

        # Write output for this respondent to JSON file
        dict_out = dict(series)
        dict_out['index'] = index
        dict_out['responses'] = dict_responses
        file_out.write_text(json.dumps(dict_out))
    except:
        # Overflow usually takes 1-2 minutes to resolve
        sleep(60)
```

`anes_simp.csv` is a CSV file containing demographic information for each synthetic respondent. A typical entry, formatted as a Pandas series, looks like:

```
year                                 2016
raceth                non-Hispanic white
age                                    29
gender                               male
ideo                    a conservative
PID                          Republican
```

```
income                                   $80,000
regis                                 registered
education                  a high school diploma
interest                              frequently
marst                                    married
```

Python dependencies (`requirements.txt`) are as follows:

```
annotated-types==0.5.0
certifi==2023.5.7
charset-normalizer==3.2.0
filelock==3.12.2
fsspec==2023.6.0
huggingface-hub==0.16.4
idna==3.4
kaleido==0.2.1
nodeenv==1.8.0
numpy==1.25.1
packaging==23.1
pandas==2.0.3
patsy==0.5.3
plotly==5.15.0
pydantic==2.0.3
pydantic_core==2.3.0
pyright==1.1.317
python-dateutil==2.8.2
pytz==2023.3
PyYAML==6.0
requests==2.31.0
scipy==1.11.1
six==1.16.0
statsmodels==0.14.0
tenacity==8.2.2
tqdm==4.65.0
typing_extensions==4.7.1
tzdata==2023.3
urllib3==2.0.3
```

14.2 ANES in the Training Data

A potential concern for comparing synthetic samples to published benchmarks is that large language models may be trained on those very benchmarks. Any correspondence between synthetic results and a baseline survey might therefore be due to the model essentially regurgitating information from its training set, meaning these models might perform considerably worse at synthetic sampling tasks where published benchmarks do not exist. Though the black-box nature of pretrained LLMs means we cannot pin down precisely why they yield particular responses, with an open source model we can at least examine whether its training set includes information about our comparison benchmark.

The Falcon-40B-Instruct LLM is trained on a refinement of the Common Crawl data [Penedo et al., 2023]. From examining the public extract of the training data published on HuggingFace,[10] it appears that the earliest data is from the summer 2013 crawl (CC-MAIN-2013-20) and the latest is from September/October 2022 (CC-MAIN-2022-40).[11] As a first cut to examine the possibility of our benchmark data appearing in the Falcon training data, we obtain the set of pages from `electionstudies.org` that appear in the September/October 2022 Common Crawl corpus. We use the `cdx-index-client.py` tool[12] to query the Common Crawl Index API[13] for pages matching the `electionstudies.org` domain. Specifically, we use the command

```
$ python cdx-index-client.py -c CC-MAIN-2022-40 "electionstudies.org/*" -v -j
```

which runs the API call

```
https://index.commoncrawl.org:443
"GET /CC-MAIN-2022-40-index?url=electionstudies.org%2F%2A&page=0&output=json HTTP/1.1"
```

This returns a JSON file containing metadata about pages from the ANES's website that appear in the Common Crawl data.

Overall, the September/October 2022 Common Crawl contains 482 pages from `electionstudies.org`. These consist of 355 HTML files, 115 PDFs, and 12 plain text files. The raw data files are behind a login wall and thus do not appear to be included in the Common Crawl.[14] The only pages apparently pertaining to the 2016 or 2020 time series studies we use as a benchmark are:

- https://electionstudies.org/2016-time-series-updates-errata/
- https://electionstudies.org/2020-time-series-study/2020-time-series-updates-errata/
- https://electionstudies.org/2020-time-series-study/2020-time-series-updates-errata/%22
- https://electionstudies.org/anes_timeseries_2020_methodology_report/
- https://electionstudies.org/anes_timeseries_2020_nrfu_userguidecodebook_20211118/
- https://electionstudies.org/data-center/2016-time-series-study/
- https://electionstudies.org/data-center/2020-time-series-study/2020-time-series-updates-e
- http://www.electionstudies.org/studypages/anes_timeseries_2016/anes_timeseries_2016_varlist.pdf
- https://electionstudies.org/studypages/anes_timeseries_2016/anes_timeseries_2016_varlist.pdf
- https://electionstudies.org/updates-announcements/anes-announcement-2020-time-series-prel

[10] https://huggingface.co/datasets/tiiuae/falcon-refinedweb.

[11] See https://commoncrawl.org/the-data/get-started/.

[12] https://github.com/ikreymer/cdx-index-client

[13] https://index.commoncrawl.org/

[14] As a first step to check if the raw data could have been obtained by other sources, we also queried for pages from the Harvard Dataverse. But the crawl contains only 70 pages total from Dataverse, none of which appear to be data files.

- https://electionstudies.org/wp-content/uploads/2021/11/anes_timeseries_2020_nrfu_userguidecodebook_20211118.pdf
- https://electionstudies.org/wp-content/uploads/2022/08/anes_timeseries_2020_methodology_report.pdf

None of these files appears to contain raw data or even summary statistics about the feeling thermometer scores we use in our analyses. There are publicly available pages on the ANES website containing such data (e.g., https://electionstudies.org/wp-content/uploads/2018/12/anes_timeseries_2016_userguidecodebook.pdf), but they do not appear at least in the most recent crawl employed by Falcon-40B-Instruct. Altogether, while we can confirm that the model is trained on some official documentation about the 2016 and 2020 ANES studies, there is not immediate evidence that our key outcomes of interest appear directly in the training data.

References

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

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

J. D. Clinton and C. D. Kam. An affective pandemic? Working paper, Vanderbilt University, 2022.

M. S. Levendusky and N. Malhotra. (mis) perceptions of partisan polarization in the american public. *Public Opinion Quarterly*, 80(S1):378–391, 2016.

X. Marquez. Gptdemocracyindex. Website, 2023. URL https://xmarquez.github.io/GPTDemocracyIndex/GPTDemocracyIndex.html.

G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023.

P. Y. Wu, J. A. Tucker, J. Nagler, and S. Messing. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057*, 2023.