Inference, Memory, or Stereotypes? Understanding LLM Roleplaying of Policy Opinions*

Boris Shor† Ryan Kennedy‡ Amanda Austin§

January 14, 2026

Abstract

This study evaluates the performance of 18 large language models (LLMs) in predicting individual responses and public opinion on policy issues. The models, spanning cloud-based, large, medium, small, and micro categories, are compared against real-world data from the Cooperative Congressional Election Study (CCES). Our demonstrate substantial variance in the performance of different LLMs, reasons to be dubious about the underlying reasoning used by LLMs to generate their responses, weaknesses in subgroup analysis (often cited as a strength of using synthetic respondents), and severe issues with sensitivity to prompt wording (even with the exact same content). While LLMs outperform random guessing, they are not yet reliable for practical use. Larger models, especially cloud-based ones, generally perform better, though local models excel in high-confidence responses. Analysis of the factors influencing LLM responses reveal a heavy reliance on partisan stereotypes, with much less evidence they accurately base estimates on other respondent characteristics. Contrary to some public reporting, LLMs struggle with estimating opinions of rarer subgroups in the population.Additionally, model predictions can vary dramatically based on the *order* of inputs into the prompt, even when the wording and actual information remain the same. These findings highlight the challenges of using LLMs for accurate and stable public opinion simulation and areas for technical improvement.

Keywords: key1, key2, key3

*abc
†University of Houston
‡The Ohio State University
§University of Houston

JEL Codes: key1, key2, key3

1 Introduction

We evaluate the performance of large language models (LLMs) in predicting both individual responses and broader opinion on public policies. A large and growing body of literature, both in political science and computer science, has suggested that LLMs are quite accurate at simulating public opinion (****). These scholars have argued that such tools will allow scholars to use off-the-shelf LLMs to estimate public opinion (****), act as synthetic participants in experiments (****), and even simulate leadership decision-making (****). Commercial entities have been quick to jump on these assessments, with start-ups raising millions of dollars to generate synthetic marketing surveys,[1] and larger companies, like Google, developing synthetic systems for debating ethical issues with AI.[2]

Some previous studies in political science have suggested caution in these conclusions. Bisbee et al. (****), for example, found that, while off-the-shelf LLMs can accurately reproduce top-line public opinion results, they are not able to reproduce inferences (i.e., the links between respondent characteristics and their opinions), can change significantly over time as models update, and perform worse when partisanship and ideology characteristics are omitted. Subsequent research has, however, generally cited this study as an example of how successful LLMs can be in reproducing top-line public opinion (****).

This study pushes the research much further. First, while previous studies have relied on a single LLM (usually the latest GPT model), this study uses a variety of commercial and open, cloud and local models. Providing a more comprehensive view of LLM performance in this role-playing task. Second, while previous studies have relied on a handful of issues on which to evaluate LLM performance, this study selects a much broader set of issues - allowing for evaluation across issues that vary in salience, support, and polarization. Third, it evaluates both individual-level and top-line results, giving a much clearer picture of both differences in LLM performance and how LLMs are making their inferences. Fourth, it

1

2

explicitly tests the accuracy of these models for rarer respondent categories, considered one of the main advantages of synthetic respondents. Finally, it tests LLM sensitivity to truly random variations in prompt design, as opposed to modifications in the actual information provided to the LLM.

The results demonstrate that, while LLMs outperform random guessing, they are still not sufficiently accurate for practical use by marketers or public officials. Cloud-based commercial models generally have higher accuracy, but local "open" models have the potential to deliver highly accurate predictions under certain conditions. Across all issues and models, both individual-level and top-line accuracy demonstrates error rates across issues that are likely to be unacceptable for practical use. We do not observe the scale of improvements across models over time that would suggest these issues will be resolved as LLMs continue development.

We also identify a clear pattern in the factors LLMs use to make inferences. LLMs heavily rely on the reported partisanship of respondent profiles, suggesting that the inferred stances of individuals are generally a product of stereotypes of partisan stances. This reliance on partisanship persists across all of the issues analyzed in this study.

Contrary to suppositions in previous studies and in popular media that LLMs may have greatest utility in simulating opinion for hard-to-reach groups, we find that the performance of these models deteriorates substantially for less common groups (e.g., Black Republicans).

Perhaps most concerning for scholars, truly random variation in prompt design produces substantial variation in top-line results. Providing an LLM with the *exact same information* about respondents and the *exact same wording*, but simply changing *the order* in which the information is presented, can produce double-digit changes in top-line public opinion estimates.

All of this suggests that, while off-the-shelf LLM performance is impressive when compared with the ease-of-use and performance of other sources of synthetic data, they are far from "disrupting" public opinion research in a significant manner (*****). We end by dis-

cussing recent innovations that seem to improve LLM performance in producing synthetic study respondents, but note that these suggest researchers may achieve better results by hybridizing traditional public opinion research with LLMs, rather than looking to them as a potential replacement for traditional public opinion research.

2 Synthetic public opinion

Public opinion research, a cornerstone of political science, has faced significant challenges over the past few decades (Berinsky, 2013). Growing difficulty in obtaining representative survey data has emerged as one of the most pressing issues (Plewes and Tourangeau, 2013). One major issue is the dramatic decline in response rates to traditional survey methods, such as telephone surveys, which has made it increasingly difficult to collect data with adequate sample sizes, especially for subgroup analysis (Kennedy and Hartig, 2019). This decline has exacerbated nonresponse bias, particularly when response rates vary systematically across subpopulations, such as by age, race, or political affiliation (Simmons and Hare, 2023). Nonresponse bias becomes especially problematic when it is linked to relevant but unobservable characteristics, such as political leanings or voting behavior (Groves and Peytcheva, 2008). Even surveys that appear to have large sample sizes face the issue of sparse cells—a problem in which observations become sparse when stratifying samples by multiple characteristics of respondents (Wang et al., 2015). This issue, referred to as the "curse of dimensionality," makes it difficult to draw valid inferences from subpopulations, complicating research on political behavior and public opinion (Bellman, 1957; Ornstein, 2020).

To address these issues, social scientists have turned to various statistical techniques, such as multilevel regression and poststratification (MRP) (Gelman and Little, 1997; Park et al., 2006). MRP and its variants have become the state-of-the-art methods for estimating subgroup opinion, especially when dealing with hierarchical or multilevel data, such as regional or demographic subgroups. These methods aim to borrow strength across observations

by pooling information from larger groups to make inferences about smaller, more sparsely populated subgroups. However, these techniques often rely on assumptions that may not always hold, and the "sparse cell" problem persists in many contexts (Little, 1993). In light of these challenges, researchers have explored alternative data sources and new technologies that could help generate more reliable and accurate public opinion data.

One technology that has garnered significant attention in recent years is the use of synthetic data generated by Large Language Models (LLMs). Broadly defined, synthetic data refers to data produced by computational models that replicate the characteristics of real-world data without using actual observations. LLMs, trained on large and diverse text corpora, have been explored as a tool for generating synthetic public opinion data (e.g., Argyle et al., 2023). These models learn complex statistical relationships between demographic variables and the language used in political discourse. During training, LLMs capture not only word associations but also higher-order interactions between variables, optimizing for the likelihood of token sequences (words or phrases) based on their context. This method allows the models to generalize from patterns in the data, generating survey responses intended to reflect the political opinions of different demographic groups. The generation process typically involves conditioning the model on prompts that specify the characteristics of the target respondent, such as their demographic profile or political beliefs, enabling the LLM to produce responses to survey items.

The use of synthetic data has been a fundamental aspect of the data science field for decades, evolving alongside advancements in computational techniques. Early examples of synthetic data generation include Monte Carlo simulations, which were used to generate artificial data for testing statistical methods and assessing model robustness in the absence of real data (Mooney, 1997). Another key approach has been the fusion of multiple datasets through traditional statistical techniques, such as those outlined by Little and Rubin (1989), to create synthetic data. Additionally, techniques like missing data imputation have long been employed to estimate missing values based on observed patterns within available data

(Rubin, 1976). Over time, these methods have advanced into more sophisticated machine learning approaches, such as Amelia II (Honaker et al., 2011), which handles missing data through multiple imputation. Another important application of synthetic data generation is addressing class imbalance—the unequal representation of various groups within a dataset. A range of machine learning techniques have been developed to mitigate class imbalance, ensuring synthetic data reflects balanced and representative distributions across subgroups (e.g., He et al., 2008). In addition to filling gaps in real datasets, synthetic data can fully replace actual data, particularly in the creation of synthetic populations—artificially generated groups designed to mirror real-world demographic characteristics. These synthetic populations have been widely applied in demographic modeling, behavior prediction, and policy analysis (e.g., Jiang et al., 2024; RTI, 2020; Benedetto et al., 2013).

The ability of LLMs to generate synthetic data at scale has sparked optimism among social scientists as a potential solution to the challenges of obtaining representative survey data. However, concerns remain about the accuracy of this synthetic data, with questions about whether LLMs are accurately reflecting real-world public opinion. In response, researchers have begun developing methods to assess and quantify the errors inherent in this process.

The Promises of Synthetic Public Opinion

Social scientists are excited about the potential of LLMs to generate synthetic samples, as these models can produce data without the logistical challenges associated with traditional data collection methods. LLMs facilitate the use of longer survey instruments that would normally be limited by human respondent constraints. A key advantage is their ability to handle extended surveys while preserving data quality, effectively avoiding common issues such as respondent fatigue and inattention (Bail, 2024; Messeri and Crockett, 2024).

LLMs can be remarkably human-like in their ability to mimic various aspects of human behavior and psychology. LLMs have been shown to replicate human moral judgments and

behavioral tendencies with remarkable accuracy (Dillion et al., 2023). These models align closely with human ethical decision-making, as evidenced by their ability to predict and mimic real-life moral choices (Dillion et al., 2023). In situations where the moral course of action is clear, LLMs tend to select commonsense responses, while in ambiguous situations, they express uncertainty (Scherrer et al., 2024). This alignment also extends to the prediction of social behaviors, such as trust, sociotropism, and competition (Leng and Yuan, 2023; Xie et al., 2024; Zhao et al., 2024). LLMs can even capture the general public's perceptions of public figures' personality traits, further demonstrating their versatility and accuracy in simulating human-like behaviors (Cao and Kosinski, 2024).

The promise of LLMs extends beyond social behavior prediction to the field of economics. LLMs have demonstrated the ability to replicate results from classic behavioral economics experiments involving concepts like self-interest, fairness, price gouging, federal budget allocations, minimum wage, and the ultimatum game (Aher et al., 2023; Horton, 2023). They can also generate detailed economic sentiment when exposed to historical news (Bybee, 2023). This economic sentiment not only closely matches existing survey measures but also reflects the deviations from rational expectations exhibited by humans.

LLMs have demonstrated significant potential to advance political science and public opinion research by providing new ways to simulate political behaviors and preferences. They can assess the positions of politicians on key policy matters (Wu et al., 2023) and measure public views on contentious topics, such as climate change (Lee et al., 2024). These models are also effective tools for estimating vote choice (Qi et al., 2024). Furthermore, generative agents have been shown to replicate participants' responses on the General Social Survey with high accuracy, matching how participants would answer their own questions two weeks later, including on topics like political party affiliation and ideology (Park et al., 2024). The ability of LLMs to generate synthetic samples suggests that they could be valuable for estimating public opinion, especially in areas where traditional data collection methods are constrained, such as in non-democratic regimes (but see Qi et al., 2024). They may even

be able to predict public reaction to political events that have yet to occur (Wang et al., 2024b).

The ability of LLMs to enhance public opinion research extends beyond generating synthetic data. These models can also play a supportive role in various stages of research. For instance, LLMs can pre-test new survey questions and assist in developing item scales (Bail, 2024). They can substitute for human respondents who drop out of longitudinal studies, helping to maintain sample integrity. Additionally, LLMs can annotate open-ended data collected from either human or synthetic samples with minimal supervision, streamlining the data analysis process (Ziems et al., 2024). While social scientists express optimism about the potential of LLMs to transform public opinion research, challenges remain in ensuring that the synthetic data they generate accurately reflects human public opinion.

The Pitfalls of Synthetic Public Opinion

The literature identifies several potential pitfalls associated with using LLMs to generate synthetic samples for public opinion research. These include the risk of memorization of training data, where models might reproduce specific details rather than generate new inferences, as well as sensitivity to prompt formulations, which can lead to inconsistent or biased outputs. Additionally, LLMs are sensitive to ordering effects, which calls into question the reliability of the models' output. Furthermore, variations across different LLMs can undermine the reliability of the generated samples, and the models' tendency to generalize may introduce distortions, presenting a particularly challenging issue for political science research.

A significant concern when using large language models (LLMs) to generate synthetic samples for public opinion research is their potential to memorize and reproduce text rather than make genuine inferences. During training, LLMs are exposed to vast amounts of text, portions of which they may memorize. This memorization can lead to models recalling and outputting specific text they encountered during pre-training, which can inflate performance metrics when benchmarked on datasets that overlap with the training data, a phenomenon

known as "data leakage" (Chang and Bergen, 2024). Memorization in language models can be somewhat mitigated by using de-duplicated data during pre-training or by raising the sampling temperature when generating text (Chang and Bergen, 2024). However, research suggests that LLMs are also capable of making inferences, especially when exposed to novel contexts after pre-training (Misra et al., 2023). While large models are more likely to generate memorized text (Chang and Bergen, 2024), they also appear capable of extrapolating patterns from memorized examples without overfitting, suggesting some degree of generalization (Tirumala et al., 2022). This difference between memorization and inference is critical, yet distinguishing between the two can be challenging for researchers (Simmons and Hare, 2023). It remains unclear whether LLMs' strong performance compared to major surveys like the American National Election Studies (ANES) (Argyle et al., 2023) reflects genuine generalization or is merely the result of memorization (Simmons and Hare, 2023). This uncertainty complicates the use of LLM-generated data for accurately predicting public opinion in political science research because it is difficult to determine whether synthetic samples reflect real-world public sentiment rather than artifacts from the training process.

Synthetic samples generated by large language models (LLMs) exhibit notable ordering effects. Dominguez-Olmedo et al. (2023) finds that LLMs, when evaluated using standard prompting methodologies and benchmarked against the American Community Survey, show significant biases in response ordering and labeling, such as a preference for options labeled with "A." These biases can be mitigated by randomizing the order of answer options, but when this is done, models tend to produce uniformly random responses, regardless of model size or pre-training data language. Additionally, research by Zhao et al. (2021) and Lu et al. (2022) demonstrates that LLM performance on multiple-choice questions can vary dramatically depending on the order of the few-shot examples provided in the prompt. Even slight changes in the order of training examples can lead to fluctuations in performance, ranging from random guessing to near state-of-the-art results. This instability is attributed to biases in LLMs toward certain answers, such as those placed toward the end of the prompt

or commonly encountered during pre-training. Further, Robinson and Wingate (2023) find that the choice of prompt order significantly influences multiple-choice question answering (MCQA) tasks, pointing out that this issue is not limited to specific models or datasets.

Large language models (LLMs) are also sensitive to the specific wording of prompts, which can significantly influence the outputs they generate. LLMs can be influenced by particular examples and phrasing when applying linguistic rules and world knowledge, leading to variations in responses based on subtle changes in prompt formulation (Chang and Bergen, 2024). This sensitivity can result in instances of memorization or under-generalization of observed examples (Chang and Bergen, 2024). Additionally, LLMs often struggle with prompts that use negation (Chang and Bergen, 2024). While humans are also sensitive to changes in wording on survey items (Smith, 1987; Rasinski, 1989, but see Huber and Paris 2013), the issue with LLMs is that they are not only affected by the phrasing of the survey question itself but also by the prompts used to instruct them. For public opinion research, this dual sensitivity to both survey question wording and instructional prompts presents a challenge, as even minor adjustments in either domain could lead to inconsistent or biased synthetic samples, undermining the reliability of the generated data and complicating efforts to accurately predict public opinion.

Another significant pitfall when using large language models (LLMs) to generate synthetic samples for public opinion research is the inconsistency across different models. Research has shown that different LLMs can exhibit varying combinations of traits and performance, which can lead to discrepancies in their outputs. For example, psychometric inventories, such as those used to assess the Big Five personality traits, have been repurposed to evaluate LLMs, revealing that different models demonstrate distinct psychological profiles (Pellert et al., 2024). Additionally, a study by Yang and Menczer (2023) audits LLMs from three major providers — OpenAI, Google, and Meta — and finds that these models differ in their ability to assess the credibility of news sources. They note that, in general, larger models are more likely to refuse to provide ratings due to insufficient information, while smaller

models are more prone to hallucinating ratings. Importantly for political science, differing LLM models have been found to possess their own unique ideological biases, across both the social and economic dimensions of ideology, potentially influencing the generation of politically biased content (Feng et al., 2023). These inconsistencies across models complicate the process of generating reliable synthetic samples. To address this issue, researchers either need a thorough understanding of the relative strengths and weaknesses of different models or must be able to identify high-quality models to include in an ensemble model, which could potentially mitigate individual biases and inconsistencies.

Large language models (LLMs) are prone to both overgeneralizing and undergeneralizing, which poses significant challenges for the generation of synthetic samples for public opinion research. Overgeneralization occurs when LLMs amplify certain patterns or biases, often reflecting stereotypes or skewed views that may not align with real-world data. For example, LLMs may replicate and even intensify gendered differences in opinion, such as in debates about the harms of misinformation, leading to artificial divides that do not exist in the broader population (Neumann et al., 2024). LLMs also overestimate ideological political polarization and the certainty of partisans (Bisbee et al., 2023), although humans also exhibit skewed perspectives of these traits (Blatz and Mercier, 2018). Possibly as a result of polarization and discussions about polarization in its online pretraining data, LLMs have also been shown to overestimate political homophily, meaning they exaggerate the ideological alignment within social networks compared to actual social dynamics (Chang et al., 2024). Conversely, LLMs tend to undergeneralize by failing to capture the full complexity of human identity and experiences, particularly when dealing with nuanced social dynamics. LLMs have also been found to "flatten" identity groups by misrepresenting the diversity of perspectives within them, which can lead to oversimplified portrayals of groups, reducing complex identities to fixed, stereotypical traits (Wang et al., 2024a). Moreover, when LLMs take on personas, they may manifest deep-seated biases, such as racial or gender-based stereotypes, which can affect their responses and performance on reasoning tasks (Gupta et al., 2023).

These stereotypes can be intersectional to up to four demographic groups (Ma et al., 2023). While LLM creators may attempt to reduce bias and stereotyping, de-biasing methods often fail to address the full range of these issues (Ma et al., 2023). Potentially as a result, LLMs can exhibit behaviors that mimic human biases, such as self-censorship in response to gender or racial biases (Lehr et al., 2024). To the extent that these biases and stereotypes accurately reflect human biases, they may be of use for researchers of public opinion, but to the extent that they flatten identity groups by portraying their opinions stereotypically, this practice is problematic for public opinion researchers. Given these tendencies to over and undergeneralize, researchers are urged to approach LLM-generated data with more cautious optimism (Korinek, 2023).

For political scientists, one of the most pressing issues is that LLMs' major biases appear to be overtly political in nature. These biases can distort their outputs, particularly in contexts related to public opinion. For example, several studies have documented that LLMs exhibit a liberal bias and a particular focus on U.S.-centered issues (Bang et al., 2024; Motoki et al., 2024; Qu and Wang, 2024). Bang et al. (2024) argue that while LLMs overall exhibit liberal biases, these biases are not uniform and can vary depending on the specific political issue being addressed, with models from the same family displaying different political leanings. This is further corroborated by Qu and Wang (2024), who show that LLMs systematically predict the opinions of educated, upper-class, liberal, white American males more accurately than those of other demographic groups, pointing to a disproportionate representation of certain social classes and political viewpoints. Motoki et al. (2024) further show that ChatGPT demonstrates a clear and systematic political bias toward the Democratic Party when responding to political questions. The political biases of LLMs are also evident in their assessments of news source credibility. Yang and Menczer (2023) reviewed several widely used LLMs and found that the models consistently exhibited a liberal bias in rating news outlets, especially when comparing left-leaning and right-leaning sources. These biases in credibility ratings became even more pronounced when LLMs were assigned partisan

identities. In addition, LLMs tend to show greater diversity of opinions on social issues than on economic ones, suggesting that they offer more nuanced analyses of social topics while remaining more uniform in their treatment of economic issues (Feng et al., 2023). Other studies have found that LLMs tend to reflect political biases that misalign with the views of specific demographic groups. For example, Santurkar et al. (2023) identified substantial misalignments between the views represented in synthetic samples and those of U.S. demographic groups, particularly older adults, Mormons, and widowed individuals. This suggests that LLMs fail to accurately capture the political diversity of different age, religious, and marital groups in the U.S. Furthermore, studies such as those by Wright et al. (2024) show that LLMs often rely on tropes and recurrent patterns in their political analyses, leading to oversimplified representations of political ideologies and values. Moreover, these models struggle with representing not only political opinions and ideologies but also the emotional and moral sentiments tied to political issues. He et al. (2024) found that LLMs can be fine-tuned to better capture the emotional nuances of political discussions, particularly in terms of political affect, but this gain comes at a loss of accuracy in estimating public opinion on policy issues. This combination of ideological biases and insufficient representation of diverse political perspectives presents a major challenge for political scientists seeking to use LLMs as tools for public opinion simulation. While LLMs may provide valuable insights into general political trends, they are less effective in capturing the full range of political attitudes across various subgroups.

The Measurement of Synthetic Public Opinion

In the study of synthetic public opinion, social scientists stress the importance of measuring the accuracy, biases, and uncertainty of synthetic samples, particularly when generated by large language models (LLMs). A central recommendation in this area is to benchmark these synthetic "silicon samples" against human samples to ensure that the data generated by LLMs reflects real-world human responses. Sarstedt et al. (2024) argue that such bench-

marking is essential to avoid the risk of synthetic data diverging from human reactions, which could lead to misleading conclusions. While benchmarking can negate some of the time and cost advantages of silicon sampling, it remains crucial for validating the synthetic data's reliability. Researchers are encouraged to use a streamlined approach, such as a small-scale human benchmarking study or secondary data, to compare results from various LLMs. However, Sarstedt et al. (2024) also note that it is vital to avoid overfitting the silicon sample to a single human benchmark, as doing so may yield results that align too closely with one human sample but fail to generalize to others.

Although benchmarking can be done for experimental methods or open-ended survey items (e.g., Aher et al., 2023; Benkler et al., 2023), the primary focus of this work is on methods used to benchmark closed-ended survey items. Benchmarking synthetic public opinion data for closed-ended survey items involves several specific methods that evaluate the fidelity of large language models (LLMs) in replicating human responses. One fundamental approach is to compare the distributions of responses generated by LLMs to the actual human data using standard statistical measures, such as means and the variances. For instance, Bisbee et al. (2023) note that synthetic data can exhibit biases or overconfidence, particularly in approximating mean values. To address this, they assess both the means and variances of synthetic responses compared to human data, providing insights into how well LLMs replicate not just the central tendency but the overall variability of human opinions. Moreover, Dominguez-Olmedo et al. (2023) compare the entropy of LLM-generated responses to real-world data from the American Community Survey, highlighting differences in the variability of responses across different models. Another means of assessment of the degree of agreement between synthetic and human responses on survey items is Cohen's Kappa, a statistic that adjusts for chance agreement between categorical responses (Qu and Wang, 2024). Kappa provides a measure of the consistency in responses while accounting for random chance. This measure can be complemented by others like Cramer's V, which assesses the strength of association between two categorical variables, and proportion agree-

ment, which calculates the percentage of exact matches between synthetic and real responses (Qu and Wang, 2024). Additionally, researchers can calculate correlations between continuous synthetic and human responses to assess the degree to which the synthetic data mirrors the patterns in the human data. For example, a high correlation suggests that the synthetic model is accurately capturing relationships between variables, much like Dillion et al. (2023) found with GPT-3.5 replicating human moral judgments. Additionally, the Wasserstein distance, or Earth Mover's Distance (EMD), can be used to measure the discrepancy between the probability distributions of synthetic and human responses (e.g., Sanders et al., 2023; Kaddour et al., 2023). The Wasserstein distance calculates the minimum cost of transforming one distribution into another, providing a clear quantification of how closely synthetic data matches real-world survey data. This method is particularly valuable when assessing the overall representativeness of a synthetic dataset across different survey questions and demographics. Finally, the F1 score, which balances precision and recall, is a useful method for analyzing synthetic data in imbalanced datasets. It evaluates how well the synthetic model captures both the positive (true) and negative (false) cases within each survey category, providing a measure of synthetic data quality, especially in cases where one response option dominates (Lee et al., 2024). von der Heyde et al. (2024) apply a similar approach, benchmarking a synthetic sample's performance by comparing its predicted vote choices to those from the German Longitudinal Election Study (GLES). Their evaluation includes calculating the share of matching vote choices between LLM-generated and GLES responses, along with F1 scores, to assess the accuracy of the LLM's predictions.

Some studies focus on the accuracy of synthetic data across specific subgroups within a population. For example, researchers may compare synthetic data to human responses across demographic subgroups such as age, gender, or ideology, to ensure that the synthetic sample mirrors the variation seen in real-world data. Many of the above methods, such as Cohen's Kappa, Cramer's V, Proportion Agreement, and F1 Scores, can be adapted to subgroup comparisons. In addition, calculating subgroup-specific correlations or the percentage

of matching responses between synthetic and human datasets can provide further insights into how well synthetic data mirrors real-world subgroup variation (Sanders et al., 2023).

In addition to these methods for benchmarking at the population and subpopulation levels, some studies focus specifically on the accuracy of synthetic data in replicating individual-level responses. For example, Park et al. (2024) highlight the methodological benefit of simulating specific individuals, allowing for an evaluation of how well each generative agent replicates an individual's attitudes and behaviors. They compare how accurately each agent matches an individual's survey responses, using normalized accuracy and correlations to assess performance. To account for variations in consistency, they normalize the accuracy of generative agents by participants' own replication accuracy over time. For continuous outcomes, they calculate normalized correlation, and for categorical outcomes such as survey responses, they assess accuracy by comparing agent responses to the individual's actual answers. This approach offers a more granular and individual-level assessment of synthetic data performance, complementing the broader methods used in general benchmarking.

Another important aspect of evaluating synthetic public opinion data is identifying potential biases in large language model (LLM) responses to survey questions. Several methods, drawn from established techniques in social science research, can be adapted for this purpose. One approach is to compare how the model responds to a set of control and treatment questions designed to elicit a particular bias, such as response order bias. Tjuatja et al. (2024) adopt this strategy by creating pairs of questions, where one version is a modified form of the original, designed to induce a known bias in human responses. By evaluating the response distributions between the original and modified questions, they quantify the degree of bias, calculating the change in responses to each pair. A significant deviation in response distributions indicates the presence of bias in the model's behavior, akin to human bias patterns. Furthermore, to test whether the bias is unique to the modification, Tjuatja et al. (2024) also introduce a set of "non-bias perturbations," such as typos or randomized letter changes, which humans are generally immune to. By comparing the results from these perturbations,

they can isolate whether the observed bias is specific to the question modification rather than a broader issue with the model's handling of survey responses. Additionally, to explore whether LLMs exhibit bias in a manner similar to human respondents, the study incorporates measures of model uncertainty, using normalized entropy to quantify the confidence level of the model's responses. Lower entropy indicates high confidence, while higher entropy suggests greater uncertainty. This measure allows for an examination of whether models are less susceptible to bias when they exhibit higher confidence, as humans are (Hippler and Schwarz, 1987).

A key challenge when evaluating synthetic public opinion data is quantifying uncertainty, particularly in the context of large language model (LLM) responses to survey questions. Given the nature of LLMs and the sensitivity of their outputs to prompt structures, the corresponding estimates are often associated with significant uncertainty, reflecting a wide range of plausible values, which complicates the interpretation of results. Sarstedt et al. (2024) emphasize that uncertainty in LLM outputs can be substantial, complicating the interpretation of data. This variability is akin to the stochastic nature of human responses, where different results may arise from the same prompt due to underlying noise (Demszky et al., 2023). Researchers urge social scientists to adopt methods for incorporating uncertainty intervals, similar to those used by metrologists in the physical sciences, by quantifying uncertainty more rigorously, including non-statistical components, to ensure more reliable and replicable findings. In the case of LLMs, one method proposed to acknowledge and incorporate this uncertainty is to generate multiple responses for any given prompt, allowing researchers to capture the variability in outputs and publish both the parameters and results to aid reproducibility and generalizability (Demszky et al., 2023). Another proposed method involves querying the LLM itself about its confidence in its answers. Kadavath et al. (2022) studies whether language models can evaluate the validity of their own claims and predict which questions they can answer correctly. The research suggests that larger models can be well-calibrated and perform better when predicting their own accuracy, as they can

estimate the probability that their answers are correct based on a diverse set of tasks. Such self-evaluation techniques, including the prediction of the probability that a model "knows" the answer, could play a crucial role in assessing the reliability of LLM-generated data, particularly when used in sensitive applications like public opinion surveys. The use of this technique may enhance the calibration and confidence assessment of LLMs, especially for subsets of responses where the model exhibits high confidence.

Researchers are actively developing methods to quantify uncertainty and bias in large language model (LLM) outputs, particularly concerning aspects like prompt sensitivity and stereotyping. Kim et al. (2024) emphasize the iterative process developers use to refine LLM-generated outputs by evaluating them against user-defined criteria. Their system, EvalLM, supports the refinement of prompts by providing an interactive evaluation of multiple outputs, enabling developers to improve prompts more efficiently. Additionally, researchers are working to gauge the extent of stereotypes and biased outputs in LLM responses to open-ended questions, particularly when simulating specific demographics. Cheng et al. (2023) develop a framework designed to evaluate the caricatured nature of LLM simulations, CoMPosT. CoMPosT uses statistical methods like quantitative analysis of response variation and comparative evaluation to measure the degree of caricature in LLM simulations. This measurement assesses dimensions like individuation (how well the simulation captures individual variability) and exaggeration (how much certain traits are amplified). These efforts represent a promising start, but more work is needed to effectively quantify the uncertainty around estimates of public opinion, particularly in relation to responses to closed-ended survey items.

3 How LLMs roleplay poll participants? Three possibilities.

4 Data

We focus on the 30 close-ended issue questions posed in the 2021 Cooperative Election Study (formerly the Cooperative Congressional Election Study, CCES). We choose to explore issue attitudes, since these are the most common and most influential target of political polls Morris (2022). There is substantial observational and experimental evidence that such issue polls impact political decision-making Burstein (2003); Butler et al. (2011); Wlezien and Soroka (2016); Morris (2022). As such, these types of polling answers are the ones most likely to be sought from synthetic respondents. Of the sources available, we chose the CCES for its large number of respondents (over 17,000) and its consistency with regards to asking about a range of issues. Finally, we focused on 2021 as the last year from which we had reasonable certainty that all the LLMs we had collected data.[3]

5 What do LLMs get right (and wrong)? More than we have a right to expect, but far less than optimists suggest.

We start this exploration where some studies end – asking how accurate synthetic roleplayers are at mimicking members of the public? In pursuing this question, it is essential to establish clear expectations. Given that LLMs are trained on vast corpora of internet text, which seldom includes explicit demographic data about the authors, expectations for high-fidelity,

[3]It is important to note that the training data for LLMs usually lag at least a couple of years behind the LLM's publication. Real-time updating of LLM weights, at the time of this writing, is too resource intensive to be practical.

individual-level accuracy might be modest. From this starting point, the performance reported in some prior studies appears quite remarkable.

However, some proponents have suggested that synthetic opinion could go well beyond this, and that synthetic public opinion could potentially substitute substitute for traditional public opinion research , implying a much higher standard of accuracy is required for LLMs to be reliable tools for researchers, marketers, and public officials. Whether this optimism is warranted remains an open empirical question.

This section evaluates the accuracy of LLMs at both the individual and aggregate levels, introducing several performance metrics for this purpose. We test a range of models to assess how performance correlates with increasing model capabilities. The findings indicate that while LLMs outperform random chance in predicting individual responses and aggregate public opinion, they do not yet meet the reliability standards required for professional applications. Furthermore, while larger models generally perform better, this trend is not sufficiently linear to project future capabilities with confidence.

5.1 Accuracy for Individual Respondents

We begin by assessing the accuracy of LLMs for individual respondents. This analysis is quantified using three primary metrics: Proportionate Reduction in Error (PRE), accuracy (or percent correctly classified), and Pearson's correlation coefficient. All metrics compare model outputs to the true opinions of CCES respondents.

Accuracy, or percent correctly classified (PCC), is the most intuitive metric. It measures the proportion of instances where the model's simulated opinion matches the respondent's true opinion. It is calculated as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

While straightforward, accuracy can be a misleading metric, particularly for issues with

a strong majority consensus. For example, if 90% of respondents hold the same view on an issue, a model can achieve 90% accuracy by simply defaulting to the majority position, without demonstrating any deeper understanding of opinion formation. In this way, accuracy gives equal credit for correctly predicting these 'easy' cases as it does for correctly predicting 'hard' cases where public opinion is evenly split.

Proportionate Reduction in Error (PRE) is a more stringent metric designed to address this limitation. PRE evaluates the model's performance improvement over a simple baseline of always predicting the majority opinion. It assesses whether the model has learned underlying patterns beyond the majority tendency. The formula is:

$$\text{PRE} = \frac{E_1 - E_2}{E_1}$$

where E_1 is the number of errors from the baseline prediction (predicting the majority opinion for all cases) and E_2 is the number of errors made by the model. A PRE of 0 indicates the model performs no better than the baseline, while a PRE of 1 signifies perfect prediction. Negative values are also possible, indicating performance worse than the baseline. Because it accounts for the baseline accuracy, PRE is a more robust indicator of a model's predictive power, especially on imbalanced datasets.

PRE is a more stringent measure than accuracy because it accounts for the baseline accuracy achievable by simply predicting the majority class. This makes it a harder hurdle for a model to clear, as it must not only be accurate but also demonstrate an improvement over a naive baseline. PRE is particularly useful when dealing with imbalanced datasets, where the majority class is significantly more frequent than the minority class. In such cases, a model could achieve high accuracy by simply predicting the majority class most of the time, but it would have a low PRE. By measuring the model's improvement over the majority-opinion baseline, PRE provides a more rigorous evaluation of whether the model has learned the underlying patterns of opinion, especially for more contentious issues.

Correlation measures the linear relationship between the model's simulated opinions and the true opinions of the CCES respondents. We use Pearson's correlation coefficient to assess if the model's predictions are directionally consistent with true opinions. This provides a view of linear association rather than absolute agreement. However, correlation does not measure classification accuracy and can be sensitive to outliers or non-linear relationships between simulated and human opinion.

These three metrics were chosen to provide a comprehensive assessment of model performance. While accuracy offers an intuitive measure of agreement, PRE provides a more rigorous test of predictive learning, and correlation assesses directional consistency.

Aggregating the mean metrics by model class (Table 1) reveals a general trend: performance tends to decline with model size. On average, cloud-based models perform comparably to large local models, though this varies by sample. Smaller local models demonstrate markedly lower performance.

Model Class	Sample	PRE	PCC	Cor
cloud	All	0.314	0.707	0.414
large	All	0.301	0.701	0.406
medium	All	0.284	0.695	0.383
micro	All	0.233	0.672	0.335
small	All	0.215	0.665	0.324
large	High Confidence	0.480	0.777	0.548
medium	High Confidence	0.452	0.772	0.532
cloud	High Confidence	0.452	0.768	0.528
micro	High Confidence	0.347	0.731	0.442
small	High Confidence	0.312	0.713	0.409

Table 1: Class-level metrics, aggregated by sample type (all responses or only those the model deems high confidence.

An examination of individual models (Table 2) reveals significant performance heterogeneity within each class. Among cloud-based models, Claude exhibits the highest performance, Gemini 2.0 Flash and Grok 4.1 Fast do quite well, while the various OpenAI models (GPT 5 mini, GPT 4.1 nano, and GPT 3.5) perform at a level comparable to an average large-sized local model. Within the local models, Mistral Small (22B) performs nearly as

well as Claude and surpasses Gemini 2.0 Flash, though the latter offers significant cost advantages. Notably, some medium-sized local models, such as Microsoft's Phi-4 (14B) and Olmo-2 (13B), perform very well and can be run on more modest local computing hardware than the largest models, which require substantial GPU resources.

Model	Class	Provider	PRE	PCC	Cor	Conf Hi	Conf Mid	Conf Lo
claude-3-haiku	Cloud	Anthropic	0.346	0.721	0.434	67.8	31.2	1.0
claude	Cloud	Anthropic	0.346	0.721	0.434	67.7	31.2	1.0
mistral-small.22b	Large	Mistral AI	0.339	0.718	0.416	56.1	43.4	0.5
gemini-2.0-flash	Cloud	Google	0.334	0.716	0.428	61.6	37.2	1.2
grok-4.1-fast	Cloud	xAI	0.321	0.710	0.408	88.8	11.2	0.1
mistral-small.24b	Large	Mistral AI	0.319	0.709	0.409	60.4	39.1	0.5
phi4.14b	Medium	Microsoft	0.313	0.707	0.410	61.5	38.5	0.1
olmo2.13b	Medium	Allen Inst	0.309	0.705	0.388	49.5	46.0	4.6
gpt-5-mini	Cloud	OpenAI	0.309	0.705	0.402	89.1	10.9	0.0
phi3.14b	Medium	Microsoft	0.308	0.705	0.404	71.1	28.6	0.3
gemma2.27b	Large	Google	0.303	0.702	0.409	59.5	37.9	2.6
gpt-4.1-nano	Cloud	OpenAI	0.300	0.701	0.413	43.1	55.8	1.1
granite3.2.8b	Small	IBM	0.297	0.700	0.423	89.6	10.3	0.1
gemini-2.0-flash-lite	Cloud	Google	0.294	0.699	0.412	47.6	45.9	6.5
gpt-3.5-turbo	Large	OpenAI	0.294	0.699	0.399	46.5	36.1	17.4
gpt-oss.20b	Large	OpenAI	0.291	0.697	0.407	68.6	30.0	1.4
olmo-3.1.32b	Large	Allen Inst	0.290	0.697	0.406	51.3	46.1	2.6
gemma3.12b	Medium	Google	0.281	0.693	0.387	33.9	62.0	4.1
granite4.small-h	Large	IBM	0.280	0.692	0.385	67.4	32.5	0.1
gemma3n.e4b	Micro	Google	0.276	0.691	0.366	56.9	43.1	0.0
exaone3.5.32b	Large	LG	0.265	0.686	0.399	40.2	59.3	0.5
llama3.1.8b	Small	Meta	0.255	0.682	0.339	62.4	36.9	0.7
exaone3.5.7.8b	Small	LG	0.244	0.677	0.343	71.8	27.9	0.4
gemma2.9b	Small	Google	0.240	0.675	0.372	69.0	28.1	2.9
gemma3.4b	Micro	Google	0.228	0.670	0.321	79.2	20.3	0.5
llama3.2.3b	Micro	Meta	0.218	0.666	0.322	67.8	16.4	15.7
mistral-nemo.12b	Medium	Mistral AI	0.211	0.663	0.328	60.7	35.7	3.6
exaone3.5.2.4b	Micro	LG	0.208	0.662	0.331	46.9	51.4	1.7
tulu3.8b	Small	Allen Inst	0.183	0.651	0.268	72.5	27.5	0.0
granite4.tiny-h	Small	IBM	0.074	0.605	0.197	76.5	22.6	0.8

Table 2: Model-level metrics, aggregating all issues and respondents. PRE is proportionate reduction in error, PCC is percent correctly classified, Cor is correlation. The Conf columns describe the portion that the model classified into confidence levels.

Restricting the analysis to high-confidence responses alters the performance landscape

(Table 3). While all models show improved accuracy in this subset, the relative rankings change. Large local models, particularly Exa One (32B), achieve the highest PRE, outperforming the cloud-based models. This suggests that while cloud models may have broader overall accuracy, large local models can be exceptionally precise when they register high confidence in a prediction.

Model	Class	PRE	PCC	Cor
exaone3.5.32b	large	0.566	0.813	0.623
gemma3.12b	medium	0.561	0.826	0.632
gpt-4.1-nano	cloud	0.551	0.810	0.620
gemini-2.0-flash-lite	cloud	0.549	0.793	0.589
mistral-small.22b	large	0.516	0.796	0.578
olmo-3.1.32b	large	0.513	0.786	0.574
gemini-2.0-flash	cloud	0.497	0.771	0.546
mistral-small.24b	large	0.491	0.777	0.551
phi4.14b	medium	0.490	0.780	0.556
claude	cloud	0.478	0.781	0.550
gemma2.27b	large	0.465	0.764	0.526
gpt-oss.20b	large	0.433	0.759	0.514
olmo2.13b	medium	0.431	0.780	0.525
phi3.14b	medium	0.423	0.750	0.504
exaone3.5.2.4b	micro	0.403	0.773	0.527
gemma3n.e4b	micro	0.402	0.741	0.471
llama3.1.8b	small	0.393	0.757	0.485
granite4.small-h	large	0.385	0.745	0.482
gemma2.9b	small	0.375	0.724	0.455
gpt	cloud	0.374	0.784	0.509
grok-4.1-fast	cloud	0.359	0.725	0.440
mistral-nemo.12b	medium	0.355	0.726	0.441
exaone3.5.7.8b	small	0.347	0.730	0.434
granite3.2.8b	small	0.347	0.720	0.457
gpt-5-mini	cloud	0.347	0.718	0.441
llama3.2.3b	micro	0.300	0.723	0.404
gemma3.4b	micro	0.282	0.687	0.366
tulu3.8b	small	0.272	0.705	0.370
granite4.tiny-h	small	0.139	0.639	0.253

Table 3: Model-level metrics, aggregating all issues and respondents, for high confidence responses only. PRE is proportionate reduction in error, PCC is percent correctly classified, Cor is correlation.

5.2 Population Accuracy

There are those who argue, however, that individual-level accuracy is not the primary concern. If LLMs make mistakes, but do so in such a way as to generate random noise, they could still be accurate in aggregate Argyle et al. (2023).

5.2.1 Aggregate Accuracy: Proportions

One method to test aggregate accuracy is by looking at the proportion indicating that they support or oppose a policy in the synthetic sample.

Model	Class	Mean	Median	SD	Max	Min	Majority
gemma2.27b	Large	0.105	0.065	0.104	0.483	0.000	0.733
grok-4.1-fast	Cloud	0.095	0.065	0.104	0.383	0.000	0.800
gemini-2.0-flash	Cloud	0.082	0.067	0.072	0.304	0.004	0.833
gemini-3-flash	Cloud	0.090	0.067	0.075	0.267	0.000	0.867
mistral-small.24b	Large	0.104	0.071	0.135	0.703	0.001	0.833
phi3.14b	Medium	0.114	0.081	0.112	0.561	0.010	0.800
claude-3-haiku	Cloud	0.102	0.084	0.087	0.276	0.002	0.833
granite4.small	Large	0.133	0.087	0.151	0.614	0.000	0.933
claude	Cloud	0.103	0.087	0.088	0.277	0.000	0.833
gpt-3.5-turbo	Large	0.102	0.089	0.085	0.311	0.006	0.800
phi4.14b	Medium	0.124	0.091	0.123	0.658	0.007	0.733
gemma3.4b	Micro	0.177	0.097	0.174	0.628	0.013	0.733
gpt-5-mini	Cloud	0.140	0.097	0.127	0.612	0.000	0.900
gemma3.12b	Medium	0.141	0.102	0.131	0.698	0.005	0.667
gpt-4.1-nano	Cloud	0.135	0.104	0.127	0.444	0.001	0.833
gemini-2.0-flash-lite	Cloud	0.138	0.106	0.124	0.701	0.018	0.667
exaone3.5.7.8b	Small	0.127	0.107	0.082	0.304	0.016	0.767
olmo-3.1.32b	Large	0.137	0.107	0.144	0.697	0.001	0.867
olmo2.13b	Medium	0.151	0.115	0.149	0.627	0.000	0.833
gpt-oss.20b	Large	0.127	0.116	0.097	0.440	0.007	0.833
exaone3.5.32b	Large	0.158	0.121	0.140	0.704	0.010	0.700
gemma3n.e4b	Micro	0.139	0.122	0.128	0.571	0.001	0.767
granite3.2.8b	Small	0.146	0.126	0.119	0.451	0.001	0.800
mistral-small.22b	Large	0.151	0.126	0.107	0.528	0.028	0.900
llama3.1.8b	Small	0.154	0.127	0.111	0.389	0.017	0.767
exaone3.5.2.4b	Micro	0.193	0.147	0.162	0.492	0.004	0.900
gemma2.9b	Small	0.166	0.158	0.119	0.688	0.010	0.633
llama3.2.3b	Micro	0.209	0.202	0.142	0.610	0.004	0.833
mistral-nemo.12b	Medium	0.235	0.208	0.184	0.636	0.004	0.700
tulu3.8b	Small	0.241	0.222	0.136	0.565	0.055	0.667
granite4.tiny	Large	0.222	0.228	0.149	0.575	0.006	0.667

Table 4: Sample-level error metrics for analyzed LLMs. Table includes several measures of error at the issue-level (i.e., difference between CCES levels of support for the 30 policy issues polled and the synthetic levels of support estimated by the LLMs). Synthetic estimates are generated at the individual respondent level and aggregated.

Model	Class	Democrat	Independent	Republican	Weighted Median
gpt-4.1-nano	Cloud	0.084	0.147	0.242	0.145
gemma2.27b	Large	0.096	0.105	0.271	0.146
exaone3.5.7.8b	Small	0.102	0.140	0.263	0.157
gpt-3.5-turbo	Large	0.104	0.112	0.292	0.157
gemma3.4b	Micro	0.094	0.111	0.313	0.159
gpt-5-mini	Cloud	0.109	0.207	0.190	0.160
phi3.14b	Medium	0.111	0.157	0.251	0.163
gemini-2.0-flash	Cloud	0.118	0.152	0.248	0.163
claude-3-haiku	Cloud	0.124	0.157	0.235	0.164
claude	Cloud	0.125	0.159	0.235	0.165
gemma3n.e4b	Micro	0.118	0.196	0.207	0.165
grok-4.1-fast	Cloud	0.117	0.112	0.300	0.165
gpt-oss.20b	Large	0.118	0.124	0.295	0.168
gemma3.12b	Medium	0.120	0.125	0.297	0.170
olmo-3.1.32b	Large	0.130	0.147	0.270	0.173
olmo2.13b	Medium	0.121	0.200	0.232	0.174
mistral-small.24b	Large	0.131	0.142	0.288	0.177
llama3.1.8b	Small	0.118	0.212	0.239	0.179
granite4.small	Large	0.109	0.201	0.274	0.181
phi4.14b	Medium	0.133	0.172	0.285	0.186
granite3.2.8b	Small	0.107	0.216	0.282	0.187
mistral-small.22b	Large	0.132	0.227	0.240	0.189
gemini-2.0-flash-lite	Cloud	0.130	0.174	0.308	0.191
exaone3.5.32b	Large	0.120	0.164	0.338	0.192
tulu3.8b	Small	0.138	0.265	0.280	0.214
gemma2.9b	Small	0.131	0.238	0.323	0.215
granite4.tiny	Large	0.208	0.192	0.252	0.215
exaone3.5.2.4b	Micro	0.144	0.253	0.292	0.216
gemini-3-flash	Cloud	0.200	0.136	0.333	0.217
mistral-nemo.12b	Medium	0.142	0.256	0.320	0.224
llama3.2.3b	Micro	0.120	0.303	0.331	0.231

Table 5: Median error by party subgroup for analyzed LLMs. Weighted median error calculated using party sample sizes as weights.

5.2.2 Aggregate Accuracy: Ideal Points

A more systemic attempt to look at aggregate accuracy is to use latent measurement models to assess a respondent-level signal amidst random noise using all responses simultaneously.

We estimate respondent-level ideal points using a standard one-dimensional IRT model for the human sample and the silicon sample simultaneously model-by-model. Then we use three metrics to assess LLM performance: mean squared deviation, mean absolute deviation, and correlation. The first one, our preferred measure, penalizes larger deviations significantly more than smaller ones.

Unlike the response-level metrics, here local models do particularly well. Llama 3.1 (8B) from Meta is the clear leader. It is followed by Tulu3 (8B), Exa One 3.5 (7.8B), and even the small Llama 3.2 (3B). Only after that do the cloud models show up, headed by GPT 4.1 Nano.

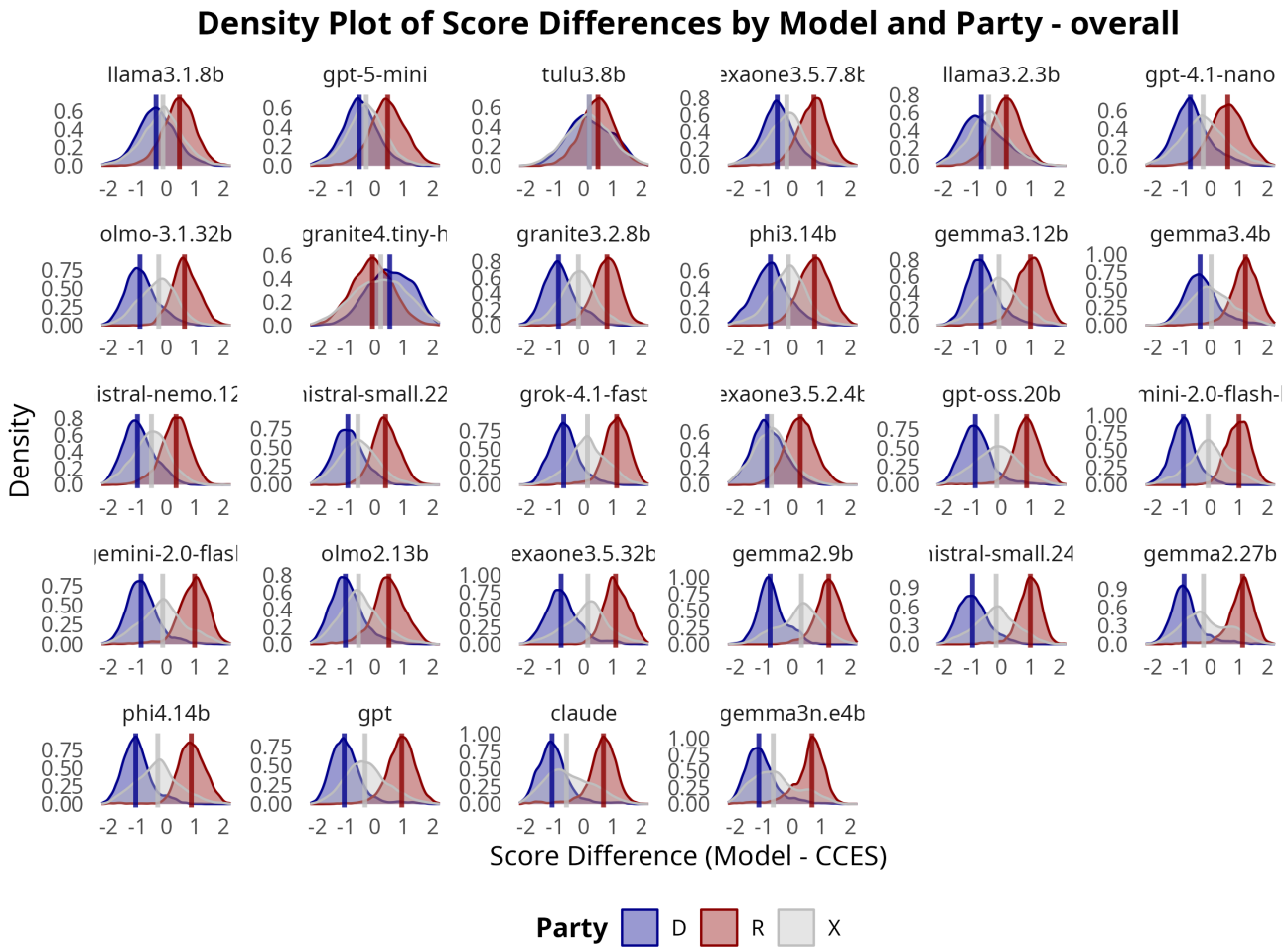

Figure 1: Estimated ideal point comparison for LLM and true respondent opinion at the model level, for all respondents at the party level.

Model	Class	Mean Sq Deviation	Mean Abs Diff	Correlation
llama3.1.8b	Small	0.531	0.581	0.748
gpt-5-mini	Cloud	0.599	0.625	0.749
tulu3.8b	Small	0.615	0.630	0.736
exaone3.5.7.8b	Small	0.637	0.656	0.743
llama3.2.3b	Micro	0.693	0.673	0.712
gpt-4.1-nano	Cloud	0.740	0.719	0.732
granite4.tiny	Medium	0.778	0.713	0.736
granite3.2.8b	Small	0.783	0.753	0.733
phi3.14b	Medium	0.791	0.738	0.730
gemma3.12b	Medium	0.792	0.763	0.754
gemma3.4b	Micro	0.821	0.748	0.718
mistral-nemo.12b	Medium	0.830	0.755	0.743
mistral-small.22b	Large	0.830	0.760	0.757
olmo-3.1.32b	Large	0.845	0.779	0.749
grok-4.1-fast	Cloud	0.846	0.788	0.760
exaone3.5.2.4b	Micro	0.853	0.764	0.719
gpt-oss.20b	Large	0.914	0.834	0.737
gemini-2.0-flash-lite	Cloud	0.930	0.844	0.739
gemini-2.0-flash	Cloud	0.932	0.840	0.740
olmo2.13b	Medium	0.937	0.826	0.732
exaone3.5.32b	Large	0.948	0.841	0.737
gemma2.9b	Small	0.975	0.860	0.736
mistral-small.24b	Large	0.990	0.874	0.739
gemma2.27b	Large	1.006	0.900	0.736
phi4.14b	Medium	1.026	0.891	0.727
claude-3-haiku	Cloud	1.076	0.915	0.700
gemma3n.e4b	Micro	1.147	0.949	0.714

Table 6: Estimated ideal point comparison for LLM and true respondent opinion at the model level, for all respondents together.

Model	Class	Weighted MSD	Democrat	Republican	Independent
llama3.1.8b	Small	0.531	0.557	0.509	0.513
gpt-5-mini	Cloud	0.599	0.663	0.576	0.527
tulu3.8b	Small	0.615	0.617	0.581	0.646
exaone3.5.7.8b	Small	0.637	0.607	0.801	0.531
llama3.2.3b	Micro	0.693	0.928	0.309	0.702
gpt-4.1-nano	Cloud	0.740	0.813	0.704	0.667
granite4.tiny	Medium	0.778	0.841	0.521	0.921
granite3.2.8b	Small	0.783	0.992	0.829	0.433
phi3.14b	Medium	0.791	0.962	0.856	0.479
gemma3.12b	Medium	0.792	0.726	1.192	0.519
gemma3.4b	Micro	0.821	0.488	1.648	0.547
mistral-nemo.12b	Medium	0.830	1.210	0.361	0.702
mistral-small.22b	Large	0.830	1.141	0.366	0.799
olmo-3.1.32b	Large	0.845	1.129	0.719	0.543
grok-4.1-fast	Cloud	0.846	0.717	1.457	0.473
exaone3.5.2.4b	Micro	0.853	1.059	0.346	1.019
gpt-oss.20b	Large	0.914	1.063	1.002	0.612
gemini-2.0-flash-lite	Cloud	0.930	1.067	1.138	0.534
gemini-2.0-flash	Cloud	0.932	0.982	1.158	0.648
olmo2.13b	Medium	0.937	1.283	0.547	0.785
exaone3.5.32b	Large	0.948	0.878	1.451	0.585
gemma2.9b	Small	0.975	0.755	1.726	0.607
mistral-small.24b	Large	0.990	1.207	1.107	0.563
gemma2.27b	Large	1.006	1.009	1.355	0.678
phi4.14b	Medium	1.026	1.285	1.024	0.648
claude-3-haiku	Cloud	1.076	1.402	0.700	0.943
gemma3n.e4b	Micro	1.147	1.536	0.654	1.030

Table 7: Estimated ideal point comparison across party subgroups. Mean Squared Deviation between LLM and true respondent ideal points by party identification. Weighted MSD is weighted by subgroup sample size.

Model	Class	D %	I %	R %	Polar %
granite4.tiny	Medium	47	942	-5	-25
tulu3.8b	Small	8	438	51	22
llama3.1.8b	Small	-70	-86	53	59
llama3.2.3b	Micro	-272	-182	24	80
gpt-5-mini	Cloud	-175	-144	59	94
exaone3.5.2.4b	Micro	-578	-204	23	112
exaone3.5.7.8b	Small	-159	-159	99	120
mistral-small.22b	Large	-763	-219	41	152
gpt-4.1-nano	Cloud	-291	-164	96	152
mistral-nemo.12b	Medium	-654	-135	54	154
phi3.14b	Medium	-321	-102	130	188
olmo2.13b	Medium	-924	-238	61	190
granite3.2.8b	Small	-394	-126	133	207
gemma3.4b	Micro	-124	-23176	289	208
gemma3.12b	Medium	-247	-181	200	217
olmo-3.1.32b	Large	-599	-137	114	222
gpt-oss.20b	Large	-475	-106	150	239
claude-3-haiku	Cloud	-1263	-234	111	256
gemini-2.0-flash	Cloud	-399	-129	212	269
grok-4.1-fast	Cloud	-258	120	279	271
gemini-2.0-flash-lite	Cloud	-456	-102	200	274
phi4.14b	Medium	-830	-126	141	274
gemma3n.e4b	Micro	-3046	-247	112	274
exaone3.5.32b	Large	-327	177	264	289
gemma2.9b	Small	-260	11789	345	307
mistral-small.24b	Large	-702	-115	198	315
gemma2.27b	Large	-486	-249	255	324

Table 8: Ideal Point Deviations by Model

Class	N	D Med	R Med	I Med	Polar Med
Small	5	-159	99	-86	120
Micro	4	-425	68	-225	160
Medium	6	-488	96	-131	189
Cloud	6	-345	156	-137	262
Large	6	-543	174	-126	264

Table 9: Ideal Point Deviations by Model Class

6 How do synthetic respondents evaluate? Radical Republicans and political polarization.

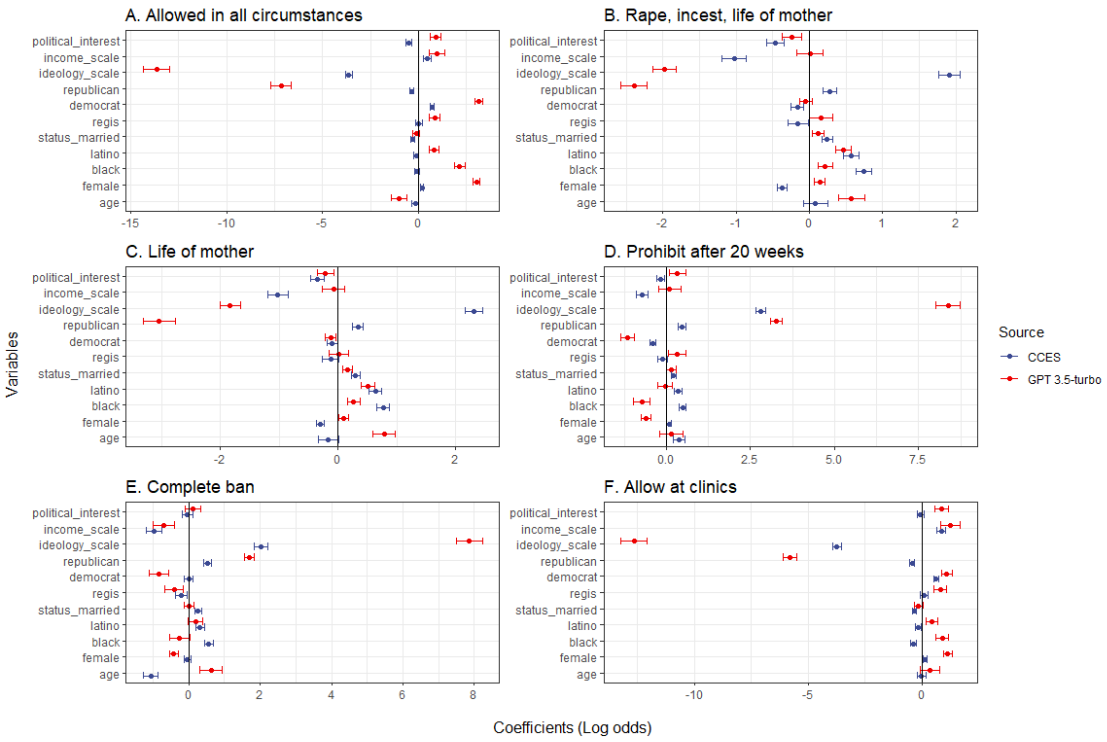

Figure 2: **Comparison of regression weights between CCES responses and GPT 3.5-Turbo generated responses for abortion issue questions.** Points show the average estimated log-odds coefficients estimated using logistic regression where 1 indicates support and 0 indicates opposition. 95% confidence intervals are calculated using 1,000 simulated draws from the posterior distribution of the regression equation. All independent variables are scaled to range between 0 and 1 for easy comparison across their ranges. Similar charts for all other issues and LLMs used in the study are available in the online appendix.

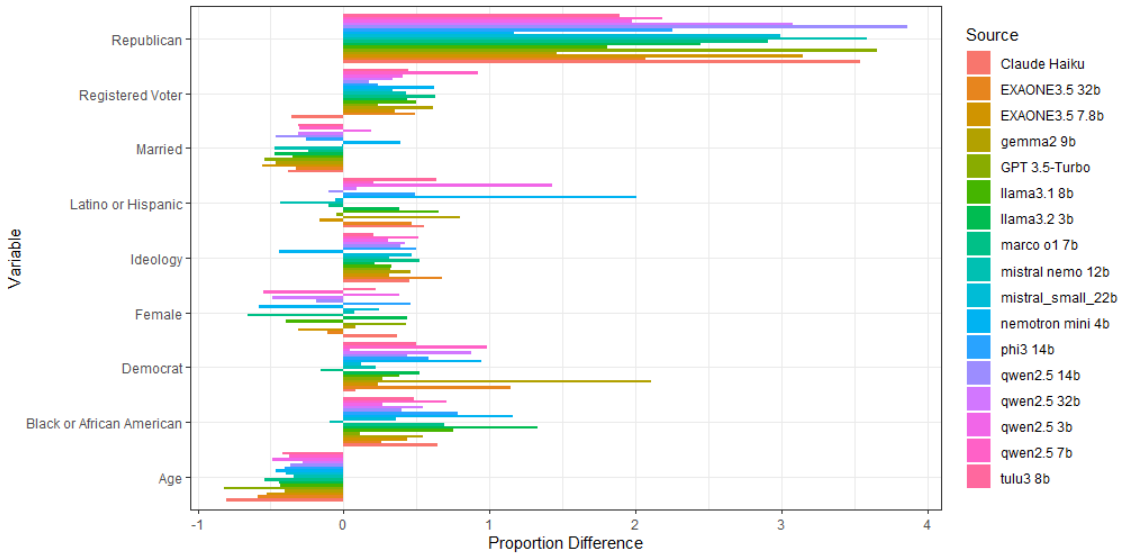

Figure 3: **Proportional difference in regression weights between LLM generated responses and CCES responses for all issue questions.** Bars show the proportional difference between the average weights across all issue models of the LLM-generated responses versus the CCES responses. Weights generated using OLS, to avoid issues of complete separation, and LLM-generated opinions that produced consensus ¿ 90% excluded to avoid convergence issues.

7 What do American voters look like to LLMs? Homogeneity or extremely extreme polarization.

8 Can LLMs model rare profiles? No...just no.

Source	mean error	median error	sd error	max error	min error	majority	N
All Respondents	0.103	0.091	0.086	0.313	0.005	0.8	17673
Democrats	0.176	0.106	0.148	0.536	0.023	0.9	7701
Independents	0.167	0.111	0.115	0.436	0.037	0.7	5168
Republicans	0.295	0.29	0.142	0.567	0.045	0.8	4804
Female Democrats	0.182	0.109	0.158	0.572	0.003	0.9	4761
White Democrats	0.165	0.091	0.147	0.512	0.016	0.9	4652
White Republicans	0.29	0.288	0.143	0.575	0.041	0.833	4156
White Independents	0.165	0.141	0.111	0.415	0.028	0.667	3803
Male Democrats	0.167	0.107	0.135	0.491	0.026	0.867	2876
Male Independents	0.178	0.184	0.106	0.439	0.004	0.667	2715
Female Republicans	0.311	0.3	0.147	0.564	0.049	0.733	2530
Female Independents	0.181	0.146	0.14	0.503	0.003	0.833	2409
Male Republicans	0.276	0.263	0.145	0.623	0.042	0.867	2268
Black & African American Democrats	0.208	0.155	0.175	0.614	0.002	0.9	1682
Black & African American Independents	0.189	0.148	0.144	0.478	0.01	0.833	514
Black & African American Republicans	0.269	0.273	0.161	0.644	0.006	0.5	121

Figure 4: **Subsample-level error metrics for GPT 3.5-Turbo.** Table includes several measures of error at the issue-level (i.e., difference between CCES levels of support for the 30 policy issues polled and the synthetic levels of support estimated by GPT 3.5-Turbo). Synthetic estimates are generated at the individual respondent level and aggregated.

9 How stable are LLM preferences? Not at all.

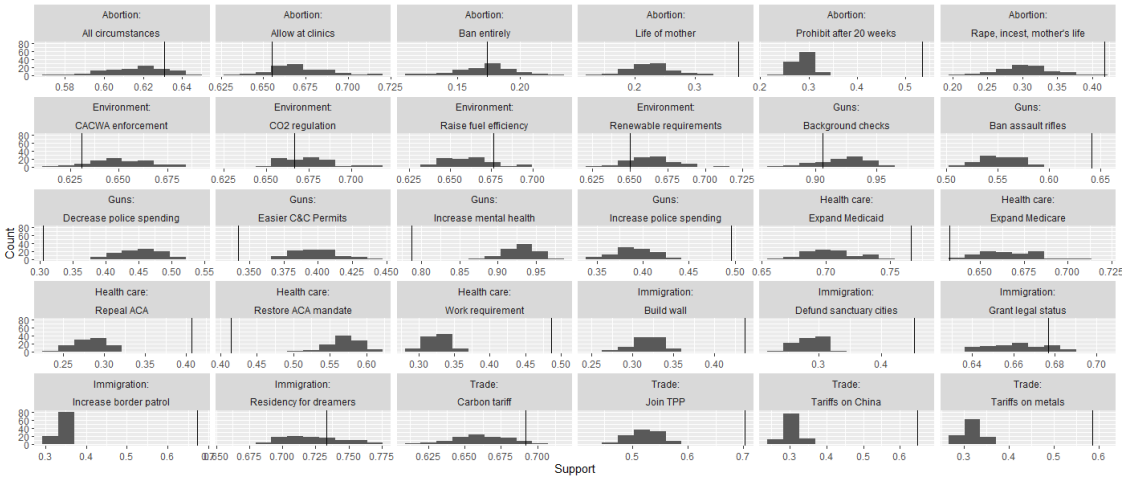

Figure 5: **Proportion supporting policies across randomly permuted entry of demographics and ideology in GPT 3.5-Turbo.** Bars show the count of runs within each bin of support proportions across 30 policy questions asked in the 2021 CCES survey. Order of demographic and ideological information presented to LLM randomized 100 times for 200 respondents.

Issue	Mean	SD	Minimum	Maximum	Range
Abortion: All circumstances	0.616	0.015	0.573	0.646	0.074
Abortion: Allow at clinics	0.671	0.016	0.633	0.718	0.085
Abortion: Ban entirely	0.171	0.023	0.112	0.226	0.114
Abortion: Life of mother	0.227	0.037	0.133	0.32	0.186
Abortion: Prohibit after 20 weeks	0.288	0.02	0.237	0.338	0.101
Abortion: Rape, incest, mother's life	0.306	0.04	0.212	0.411	0.199
Environment: CACWA enforcement	0.65	0.015	0.613	0.687	0.074
Environment: CO2 regulation	0.674	0.016	0.634	0.719	0.085
Environment: Raise fuel efficiency	0.662	0.016	0.629	0.716	0.087
Environment: Renewable requirements	0.665	0.017	0.622	0.718	0.095
Guns: Background checks	0.923	0.022	0.863	0.979	0.117
Guns: Ban assault rifles	0.554	0.019	0.503	0.594	0.091
Guns: Decrease police spending	0.45	0.034	0.362	0.523	0.161
Guns: Easier C&C Permits	0.4	0.017	0.364	0.447	0.083
Guns: Increase mental health	0.931	0.023	0.869	0.974	0.106
Guns: Increase police spending	0.391	0.021	0.34	0.447	0.107
Health care: Expand Medicaid	0.704	0.021	0.655	0.759	0.104
Health care: Expand Medicare	0.667	0.017	0.633	0.714	0.081
Health care: Repeal ACA	0.283	0.021	0.231	0.323	0.092
Health care: Restore ACA mandate	0.57	0.021	0.505	0.611	0.106
Health care: Work requirement	0.326	0.017	0.284	0.357	0.073
Immigration: Build wall	0.317	0.019	0.262	0.354	0.092
Immigration: Defund sanctuary cities	0.285	0.021	0.222	0.323	0.101
Immigration: Grant legal status	0.662	0.014	0.633	0.703	0.069
Immigration: Increase border patrol	0.346	0.013	0.321	0.381	0.059
Immigration: Residency for dreamers	0.722	0.023	0.663	0.772	0.108
Trade: Carbon tariff	0.661	0.02	0.613	0.712	0.099
Trade: Join TPP	0.521	0.028	0.444	0.584	0.139
Trade: Tariffs on China	0.306	0.019	0.256	0.343	0.087
Trade: Tariffs on metals	0.318	0.019	0.267	0.35	0.084

Table 10: **Characteristics of support proportions across information order permutations of demographic and ideological information.** Table provides statistical summaries for all 30 policy questions asked in the 2021 CCES survey. Order of demographic and ideological information randomized 100 times for 200 respondents.

10 Conclusions

11 Boris

11.1 Heatmaps

11.2 Efficiency

Model	GPU	Token Rate	Response Time	Hours for 1000
exaone3.5.2.4b	Apple M4 Pro	63	1.7	14.4
exaone3.5.32b	3090	35	2.5	21.2
exaone3.5.7.8b	4070	84	1.2	9.7
exaone3.5.7.8b	2080 Ti	77	1.3	10.6
gemma2.27b	3090	35	1.8	15.1
gemma2.9b	2080 Ti	45	1.6	13.3
gemma2.9b	4070	37	1.8	14.8
gemma3.12b	4090 Laptop GPU	41	3.8	31.5
gemma3.12b	Apple M4 Pro	27	6.7	56.0
gemma3.12b	3090	62	2.4	19.7
gemma3.4b	4090 Laptop GPU	75	2.0	16.6
gemma3.4b	4070	34	4.8	39.6
gemma3.4b	Apple M4 Pro	53	2.9	24.1
gemma3.4b	2080 Ti	83	1.8	15.3
gemma3n.e4b	Apple M4 Pro	34	3.9	32.7
gemma3n.e4b	4090 Laptop GPU	30	4.3	36.1
gemma3n.e4b	2080 Ti	49	2.8	23.2
granite3.2.8b	Apple M4 Pro	33	3.2	26.9
granite3.2.8b	4070	75	1.4	11.8
granite3.2.8b	2080 Ti	68	1.6	13.3
llama3.1.8b	4070	78	1.1	9.3
llama3.1.8b	2080 Ti	63	1.4	11.3
llama3.1.8b	3090	57	1.5	12.8
llama3.1.8b	Apple M4 Pro	27	2.8	23.4
llama3.2.3b	4070	99	0.8	6.4
llama3.2.3b	2080 Ti	70	1.2	9.7
llama3.2.3b	Apple M4 Pro	64	1.2	9.9
mistral-nemo.12b	2080 Ti	43	2.3	19.3
mistral-nemo.12b	4070	55	1.8	15.0
mistral-nemo.12b	3090	75	1.3	10.8
mistral-small.22b	2080 Ti	29	2.8	23.2
mistral-small.22b	3090	47	1.7	14.1
mistral-small.24b	3090	46	2.0	16.5
olmo2.13b	Apple M4 Pro	22	5.6	46.4
olmo2.13b	2080 Ti	46	2.6	21.6
olmo2.13b	4070	49	2.5	21.0
phi3.14b	4070	48	2.2	18.0
phi3.14b	2080 Ti	48	2.2	18.1
phi3.14b	3090	65	1.6	13.2
phi3.14b	4090 Laptop GPU	37	2.7	22.5
phi4.14b	4070	46	2.1	17.1
phi4.14b	Apple M4 Pro	19	4.8	39.8
phi4.14b	2080 Ti	45	2.1	17.3
phi4.14b	3090	67	1.4	11.8
tulu3.8b	4070	79	0.9	7.3
tulu3.8b	2080 Ti	39	1.8	15.2
tulu3.8b	3090	99	0.7	5.9
granite4.tiny-h	2080 Ti	85	1.1	9.5
granite4.small-h	3090	50	1.8	15.0

Model	GPU	PRE Efficiency
llama3.2.3b	4070	0.337
phi4.14b	3090	0.279
exaone3.5.7.8b	2080 Ti	0.244
llama3.2.3b	2080 Ti	0.240
phi3.14b	3090	0.224
tulu3.8b	3090	0.222
granite4.small-h	5090	0.218
llama3.1.8b	4070	0.218
llama3.1.8b	3090	0.210
granite3.2.8b	2080 Ti	0.210
exaone3.5.7.8b	4070	0.200
mistral-small.22b	3090	0.200
granite3.2.8b	4070	0.198
llama3.1.8b	2080 Ti	0.195
tulu3.8b	4070	0.191
gemma2.9b	4070	0.184
phi4.14b	4070	0.177
gpt-oss.20b	5090	0.174
llama3.2.3b	Apple M4 Pro	0.168
gemma2.27b	3090	0.167
phi3.14b	4090 Laptop GPU	0.161
mistral-small.24b	3090	0.161
granite4.small-h	3090	0.155
gemma3.4b	4090 Laptop GPU	0.151
gemma2.9b	2080 Ti	0.151
mistral-nemo.12b	4070	0.150
phi3.14b	2080 Ti	0.147
olmo-3.1.32b	3090	0.139
llama3.1.8b	Apple M4 Pro	0.135
phi3.14b	4070	0.132
olmo2.13b	4070	0.126
tulu3.8b	2080 Ti	0.125
mistral-small.22b	2080 Ti	0.121
exaone3.5.2.4b	Apple M4 Pro	0.120
olmo2.13b	2080 Ti	0.119
gemma3.12b	3090	0.119
mistral-nemo.12b	3090	0.117
gemma3.4b	2080 Ti	0.110
phi4.14b	2080 Ti	0.109
mistral-nemo.12b	2080 Ti	0.107
granite3.2.8b	Apple M4 Pro	0.105
exaone3.5.32b	3090	0.104
gemma3n.e4b	2080 Ti	0.099
gemma3.4b	Apple M4 Pro	0.099
phi4.14b	Apple M4 Pro	0.088
olmo-3.1.32b	5090	0.081
gemma3.12b	4090 Laptop GPU	0.080
gemma3.12b	Apple M4 Pro	0.077
olmo2.13b	Apple M4 Pro	0.067

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Bail, C. A. (2024). Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Bang, Y., Chen, D., Lee, N., and Fung, P. (2024). Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Benedetto, G., Stinson, M., and Abowd, J. M. (2013). The creation and use of the sipp synthetic beta. *2013*.

Benkler, N., Mosaphir, D., Friedman, S., Smart, A., and Schmer-Galunder, S. (2023). Assessing llms for moral value pluralism. *arXiv preprint arXiv:2312.10075*.

Berinsky, A. J. (2013). *Silent voices: Public opinion and political participation in America*. Princeton University Press.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., and Larson, J. (2023). Artificially precise extremism: how internet-trained llms exaggerate our differences. SocArXiv Preprint.

Blatz, C. W. and Mercier, B. (2018). False polarization and false moderation: Political opponents overestimate the extremity of each other's ideologies but underestimate each other's certainty. *Social Psychological and Personality Science*, 9(5):521–529.

Burstein, P. (2003). The impact of public opinion on public policy: A review and an agenda. *Political research quarterly*, 56(1):29–40.

Butler, D. M., Nickerson, D. W., et al. (2011). Can learning constituency opinion affect how legislators vote? results from a field experiment. *Quarterly Journal of Political Science*, 6(1):55–83.

Bybee, J. L. (2023). The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*.

Cao, X. and Kosinski, M. (2024). Large language models know how the personality of public figures is perceived by the general public. *Scientific Reports*, 14(1):6735.

Chang, S., Chaszczewicz, A., Wang, E., Josifovska, M., Pierson, E., and Leskovec, J. (2024). Llms generate structurally realistic social networks but overestimate political homophily. *arXiv preprint arXiv:2408.16629*.

Chang, T. A. and Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Cheng, M., Piccardi, T., and Yang, D. (2023). Compost: Characterizing and evaluating caricature in llm simulations. *arXiv preprint arXiv:2310.11501*.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., and Eichstaedt, J. C. e. a. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. (2023). Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

Feng, S., Park, C. Y., Liu, Y., and Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Gelman, A. and Little, T. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23:127.

Groves, R. M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly*, 72(2):167–189.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., and Khot, T. (2023). Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. pages 1322–1328.

He, Z., Guo, S., Rao, A., and Lerman, K. (2024). Whose emotions and moral sentiments do language models reflect? *arXiv preprint arXiv:2402.11114*.

Hippler, H.-J. and Schwarz, N. (1987). Response effects in surveys. In *Social information processing and survey methodology*, pages 102–122. Springer.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia ii: A program for missing data. *Journal of Statistical Software*, 45:1–47.

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

Huber, G. and Paris, C. (2013). Assessing the programmatic equivalence assumption in question wording experiments: Understanding why americans like assistance to the poor more than welfare. *Public Opinion Quarterly*, 77(1):385–397.

Jiang, N., Yin, F., Wang, B., and Crooks, A. T. (2024). A large-scale geographically explicit synthetic population with social networks for the united states. *Scientific Data*, 11(1):1204.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., and et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Kennedy, C. and Hartig, H. (2019). Response rates in telephone surveys have resumed their decline.

Kim, T. S., Lee, Y., Shin, J., Kim, Y.-H., and Kim, J. (2024). Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Korinek, A. (2023). Language models and cognitive automation for economic research.

Lee, S., Peng, T.-Q., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., and Leiserowitz, A. (2024). Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429.

Lehr, S. A., Caliskan, A., Liyanage, S., and Banaji, M. R. (2024). Chatgpt as research scientist: Probing gpt's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, 121(35):e2404328121.

Leng, Y. and Yuan, Y. (2023). Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*.

Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88(423):1001–1012.

Little, R. J. A. and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 8086–8098.

Ma, W., Chiang, B., Wu, T., Wang, L., and Vosoughi, S. (2023). Intersectional stereotypes in large language models: Dataset and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597.

Messeri, L. and Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

Misra, K., Rayz, J. T., and Ettinger, A. (2023). Comps: Conceptual minimal pair sentences for testing property knowledge and inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949.

Mooney, C. Z. (1997). *Monte Carlo Simulation*. Number 116. Sage.

Morris, G. E. (2022). *Strength in Numbers: How Polls Work and why We Need Them*. WW Norton & Company.

Motoki, F., Pinho Neto, V., and Rodrigues, V. (2024). More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Neumann, T., Lee, S., De-Arteaga, M., Fazelpour, S., and Lease, M. (2024). Diverse, but divisive: Llms can exaggerate gender differences in opinion related to harms of misinformation. *arXiv preprint arXiv:2401.16558*.

Ornstein, J. T. (2020). Stacked regression and poststratification. *Political Analysis*, 28(2):293–301.

Park, D. K., Gelman, A., and Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In *Public opinion in state politics*, pages 209–228.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., and Strohmaier, M. (2024). Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Plewes, T. J. and Tourangeau, R., editors (2013). *Nonresponse in social science surveys: A research agenda*.

Qi, W., Lyu, H., and Luo, J. (2024). Representation bias in political sample simulations with large language models. *arXiv preprint arXiv:2407.11409*.

Qu, Y. and Wang, J. (2024). Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Rasinski, K. A. (1989). The effect of question wording on public support for government spending. *Public Opinion Quarterly*, 53:388–400.

Robinson, J. and Wingate, D. (2023). Leveraging large language models for multiple choice question answering. In *Proceedings of the International Conference on Learning Representations*.

RTI, U. (2020). Synthetic household population.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Sanders, N. E., Ulinich, A., and Schneier, B. (2023). Demonstrations of the potential of ai-based political issue polling. *arXiv preprint arXiv:2307.04781*.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004.

Sarstedt, M., Adler, S. J., Rau, L., and Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology and Marketing*, 41(6):1254–1270.

Scherrer, N., Shi, C., Feder, A., and Blei, D. (2024). Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.

Simmons, G. and Hare, C. (2023). Large language models as subpopulation representative models: A review. *arXiv preprint arXiv:2310.17888*.

Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter: An analysis of the impact of question wording on response patterns. *Public Opinion Quarterly*, 51:75.

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 38274–38290.

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., and Neubig, G. (2024). Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

von der Heyde, L., Haensch, A.-C., and Wenz, A. (2024). Vox populi, vox ai? using language models to estimate german public opinion. *arXiv preprint arXiv:2407.08563*.

Wang, A., Morgenstern, J., and Dickerson, J. P. (2024a). Large language models should not replace human participants because they can misportray and flatten identity groups. *arXiv preprint*, abs/2402.01908.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., and Chen, Z. e. a. (2024b). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

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

Wlezien, C. and Soroka, S. N. (2016). Public opinion and public policy. In *Oxford research encyclopedia of politics*.

Wright, D., Arora, A., Borenstein, N., Yadav, S., Belongie, S., and Augenstein, I. (2024). Llm tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112.

Wu, P. Y., Nagler, J., Tucker, J. A., and Messing, S. (2023). Large language models can be used to estimate the latent positions of politicians. *arXiv preprint arXiv:2303.12057*.

Xie, C., Chen, C., Jia, F., Ye, Z., Shu, K., Bibi, A., Hu, Z., Torr, P., Ghanem, B., and Li, G. (2024). Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*.

Yang, K.-C. and Menczer, F. (2023). Accuracy and political bias of news source credibility ratings by large language models. *arXiv preprint arXiv 2304*.

Zhao, Q., Wang, J., Zhang, Y., Jin, Y., Zhu, K., Chen, H., and Xie, X. (2024). Competeai: Understanding the competition dynamics of large language model-based agents. In *Forty-first International Conference on Machine Learning*.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference on Machine Learning*, pages 12697–12706.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.