# Can Large Language Models Accurately Predict Public Opinion? A Review

Arnault Pachot, Thierry Petit

# Can Large Language Models Accurately Predict Public Opinion? A Review

Arnault Pachot
Emotia
Paris, France

Thierry Petit
Emotia
Paris, France

## Abstract

This article reviews the capacity of Large Language Models (LLMs) to accurately predict public opinion. Through a meta-analysis of recent research, we evaluate metrics such as accuracy, correlation, and bias to assess the implications for survey-based social science.

Our findings indicate that LLMs demonstrate high accuracy in predicting demographic responses and aligning with survey data, suggesting their potential in reflecting public opinion trends. Despite this, challenges persist, particularly with regard to biases in race, gender, and ideology, which can distort model outputs. Several studies propose methods to measure and mitigate these biases, aiming to improve the representativeness of LLM predictions.

By addressing these limitations and refining the ethical use of LLMs, these models could significantly enhance public opinion research, offering valuable insights for researchers, policymakers, and decision-makers in forecasting societal sentiments.

## CCS Concepts

• **Information systems** → **Social networks**; • **Computing methodologies** → *Natural language processing*; • **Theory of computation** → Machine learning theory.

## Keywords

Large Language Models (LLMs), Survey Research, Public Opinion Prediction, Subpopulation Simulation, Meta-Analysis, Accuracy, Bias, Predictive Power, Methodology, Natural Language Processing (NLP), Artificial Intelligence (AI), Data Collection, Ethical Considerations, Demographic Bias, Ideological Bias, Model Robustness, Current Events, Economic Changes, Survey Methodologies, AI-Driven Surveys

**ACM Reference Format:**
Arnault Pachot and Thierry Petit. 2018. Can Large Language Models Accurately Predict Public Opinion? A Review. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX


*Conference acronym 'XX, June 03–05, 2018, Woodstock, NY*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

The rapid progress in intelligence (AI) and natural language processing (NLP) has resulted in the creation of Large Language Models (LLMs), such as OpenAI Chat-GPT, that can emulate conversational speech, translate languages and summarize information. One of the emerging uses of LLMs is their application in survey research, providing avenues to predict opinions, model human populations and analyze survey data [8, 16].

Survey research plays a role in trends and public attitudes. While traditional survey methods are effective, they often require time and effort. They are subject to biases. The incorporation of LLMs into survey research procedures holds the promise of boosting efficiency, cutting costs and reducing those biases. Application areas include automated survey design, data collection, and context-based analysis. These advancements could greatly streamline processes making surveys more accessible and affordable, for businesses and researchers alike [13].

In this perspective, recent studies have emphasized the potential of LLMs to accurately forecast survey responses, create sample simulations, and handle datasets [8, 16, 18]. LLMs have been used to predict people opinions, to simulate how the public would react to situations, and even mirror the complexities found in human subgroups. These results indicate that LLMs can not only improve the accuracy and efficiency of opinion surveys but also open up opportunities for creating products and service-based on user preferences and trends, ultimately increasing user satisfaction. Mellon et al. [12] discovered that Large LLMs can come close, to human level accuracy when coding responses from open text surveys. However they also pointed out that there are differences in performance, among models. Recent studies have shown that LLMs can mimic behaviors and replicate human subject experiments showcasing their potential and limitations, in various research scenarios [1]. Moreover studies have demonstrated that LLMs have the ability to carry over biases from their training data to subsequent tasks impacting the equity of use cases, like identifying hate speech and misinformation [6]. Recent discoveries indicate that LLMs exhibit biases to those of humans concerning moral standards. These biases can impact their actions and results significantly (Schramowski, 2022).

This literature review aim to highlight the strengths and weaknesses of using LLMs in survey research, through a meta-analysis of existing papers. We focus on measures like accuracy, correlation and bias, while assessing how the models function under conditions influenced by social or political events, and economic changes.

Our aim is to provide insights for professionals, scholars, and decision-makers, on how to use LLMs in survey techniques efficiently and ethically. Our discoveries confirm that LLMs can bring

about in enhancing the accuracy, efficiency and impartiality of social science surveys [19]. This outcome opens up avenues for market analysis and product development aligned with user preferences.

The paper is organized as follows. The **Methodology** section introduces our criteria for choosing studies using LitMap platform and how we gathered and analyzed data. In the **Results** section, we present a in-depth and systematic comparison of performance metrics from studies. We investigate how real time updates affect model reliability. In the **Discussion** section, we examine the factors and obstacles linked to the use of LLMs in survey studies and suggest ideas for future research.

## 2 Methodology

In this section, we detail our approach to conducting the analysis. We outline our selection criteria for studies and the process of data compilation.

We selected studies that included performance measures such as accuracy, correlation, and bias indicators.

### 2.1 Search Strategy

During our search for research materials, we faced challenges with keywords because the term "survey" is commonly associated with a type of paper. We began by examining the research conducted by Argyle et al. [2]. We then used LitMap and Scholar GPT to delve into publications within the field. As this area of study continues to grow, we broadened our search to include preprint publications.

### 2.2 Data Extraction and Synthesis

Data from the selected research papers was thoughtfully integrated. Each study we included contributed valuable information for thorough examination:

- **Dataset Used**: Type and source of the dataset (e.g., ANES, GSS, GlobalOpinionQA).
- **LLMs Tested**: Specific models and their versions (e.g., GPT-3, GPT-4, Alpaca-7B).
- **LLM Parameters**: Parameters used for the LLMs (e.g., temperature).
- **Types of Prompts**: Details on the types of prompts used (e.g., demographic, ideological, economic indicators).
- **Performance Metrics**: Key performance metrics reported.

The collected data was combined to provide an overview of the techniques and results in each research study.

We categorized the publications into four themes, based on their focus and employed methodology:

- **Data and Methodology:** This theme outlines the methodologies and approaches utilized in the studies to simulate and assess opinions.
- **Performance Metrics:** This theme presents the findings and results.
- **Bias and Ethical Considerations:** This theme addresses the biais and ethical limitations and challenges identified in each paper.
- **Key Contributions:** This theme highlights the contributions and innovative aspects of each study.

## 3 Results

We have chosen a total of 10 publications. All the studies have been released within the three last years, highlighting the nature of this emerging field. They share a common focus on exploring how LLMs can replicate and predict opinions.

The articles reviewed in Table 1 offer insights into the realm of opinion polling through the use of LLMs.

All the selected studies provide an examination of the methodologies and results pertaining to LLM utilization in survey research.

A notable contribution is the concept of "algorithmic fidelity." Argyle et al. [2] demonstrate that OpenAI's Chat-GPT3 can effectively replicate sub populations by leveraging socio demographic information indicating that LLMs can act as proxies for specific human groups.

Similarly, Sun et al. [17] introduce the "silicon sampling" technique to mirror the opinions of sub groups using demographic data, shedding light on both the viability of this method and the inherent biases in LLMs, and stressing the importance of thorough bias evaluation.

Another key research direction focuses on enhancing opinion prediction precision.

Hwang et al. [7] achieve heightened accuracy by incorporating details and past opinion data resulting in LLM outputs that better align with individual user viewpoints.

Kim and Lee [9] have shown the success of an approach by achieving accuracy in predicting missing responses and unasked opinions through fine tuning LLMs with cross sectional surveys. These findings highlight the potential for LLMs to offer tailored insights based on user history and temporal data.

Durmus et al. [5] created the GlobalOpinionQA dataset to assess how well LLMs represent viewpoints globally. Their work not only identifies biases but also suggests new metrics to measure representation accuracy, contributing to a better understanding of LLM performance on a global scale.

Chu et al. [4] introduced media diet models that adjust LLMs using content from sources like news and TV broadcasts. The objective is to mirror the opinions of specific subgroups based on media consumption habits. This method was proved effective in predicting judgments across surveys, while remaining resilient to variations in wording and media exposure.

Sanders et al. [14] demonstrated the ability of AI chatbots powered by LLMs to predict public sentiment accurately, particularly in ideological contexts. They did acknowledge limitations in capturing differences, which suggests areas for enhancement.

Lee et al. [10] examined how well LLMs represent opinion on warming in specific areas. They discovered that while LLMs can mimic voting patterns accurately, they struggle to reflect perspectives on warming without detailed context. This research underscores the importance of choosing models and refining engineering processes to assess biases and enhance LLM performance in targeted fields.

Santurkar et al. [15] delved into the alignment between LLMs and demographic opinions by introducing OpinionsQA, a dataset tailored to assess how well LLMs align with the viewpoints of 60 US groups across various subjects. They uncovered discrepancies between opinions generated by LLMs and those held by groups

**Table 1: Summary of Methodologies Used in Survey Research with LLMs**

| Study | Dataset | Types of Prompts | Citations |
|---|---|---|---|
| Argyle et al. [2] | ANES 2012, 2016, 2020 (Election Studies) | Socio-demographic backstories | 271 |
| Santurkar et al. [15] | OpinionsQA (60 US demographic groups) | Demographic, Ideology | 167 |
| Durmus et al. [5] | GlobalOpinionQA (Global Attitudes) | Linguistic, Cross-national | 72 |
| Bisbee et al. [3] | ANES 2016-2020 (Election Studies) | Demographic, Political Characteristics | 39 |
| Kim and Lee [9] | GSS 1972-2021 (Social Survey) | Demographic, Temporal contexts | 24 |
| Chu et al. [4] | Nationally representative surveys (COVID-19, consumer confidence) | Media diet prompts | 19 |
| Hwang et al. [7] | OpinionQA (Survey responses) | Demographic, Ideology | 15 |
| Sanders et al. [14] | Various political opinion polls | Demographic, Issue-specific | 4 |
| Lee et al. [10] | Pew Research Global Attitudes | Demographic, Psychological covariates | 3 |
| Sun et al. [17] | Synthetic datasets (Sub-population simulation) | Random Silicon Sampling | 1 |

highlighting the necessity for better model alignment and bias mitigation strategies.

Bisbee et al. [3] critically analyzed the capabilities and constraints of using LLMs to create survey data. They pointed out hurdles in ensuring the accuracy dependability and reproducibility of data warning against substituting traditional survey methods with this approach, in social science research.

Table 1 provides an overview of the studies reviewed, detailing the datasets utilized parameters considered and prompts used in each study.

## 3.1 Argyle et al. [2] - Out of One, Many: Using Language Models to Simulate Human Samples

Authors conducted a study on the effectiveness of LLMs in replicating responses through "silicon sampling." They trained OpenAI's Chat-GPT3 on socio backgrounds to generate survey answers.

*3.1.1 Data and Methodology.* The research utilized data from the American National Election Studies (ANES) waves of 2012, 2016 and 2020 as Rothschild et al.s "Pigeonholing Partisans" dataset. GPT-3 was conditioned on socio backgrounds for producing responses. The evaluation focused on fidelity examining how well the model captures connections between concepts, attitudes and social contexts [2, Figure 4].

*3.1.2 Performance Metrics.* The performance assessment included Tetrachoric Correlation, Cohen's Kappa, Intraclass Correlation Coefficient (ICC) and Proportion Agreement.

*Tetrachoric Correlation.* This gauges the correlation between variables. The outcomes across categories are detailed in Table 2.

**Table 2: Tetrachoric Correlation between GPT-3 and ANES probability of voting for the Republican presidential candidate [2, Table 1].**

| Variable | 2012 | 2016 | 2020 |
|---|---|---|---|
| Whole sample | 0.90 | 0.92 | 0.94 |
| Men | 0.90 | 0.93 | 0.95 |
| Women | 0.91 | 0.92 | 0.94 |
| Strong partisans | 0.99 | 1.00 | 1.00 |
| Weak partisans | 0.73 | 0.71 | 0.84 |
| Leaners | 0.90 | 0.93 | 0.95 |
| Conservatives | 0.84 | 0.88 | 0.91 |
| Moderates | 0.65 | 0.76 | 0.71 |
| Liberals | 0.81 | 0.73 | 0.86 |
| Whites | 0.87 | 0.91 | 0.94 |
| Blacks | 0.71 | 0.87 | 0.81 |
| Hispanics | 0.86 | 0.93 | 0.88 |
| Attends church | 0.91 | 0.93 | 0.94 |
| Doesn't attend church | 0.88 | 0.90 | 0.93 |
| High interest in politics | 0.95 | 0.97 | 0.97 |
| Low interest in politics | 0.71 | 0.75 | 0.83 |
| Discusses politics | 0.92 | 0.94 | 0.95 |
| Doesn't discuss politics | 0.83 | 0.81 | 0.80 |
| 18 to 30 years old | 0.90 | 0.90 | 0.90 |
| 31 to 45 years old | 0.90 | 0.92 | 0.94 |
| 46 to 60 years old | 0.90 | 0.92 | 0.92 |
| Over 60 | 0.90 | 0.93 | 0.96 |

The Tetrachoric Correlation findings show connections, among categories implying that the forecasts made by GPT 3 closely match the likelihood of supporting the Republican presidential candidate as per ANES data. The strongest supporters exhibit the connection showcasing the models accuracy, in anticipating voting patterns within this segment. Conversely weaker supporters and moderates display a correlation indicating instances where the models predictions may be less accurate.

*Cohen's Kappa.* This measure evaluates agreement among raters for items while accounting for chance agreement. Specific results are outlined in Table 3.

**Table 3: Cohen's Kappa values for various groups for the years 2012, 2016, and 2020 [2, Table 8, 9, 10].**

| Variable | 2012 | 2016 | 2020 |
|---|---|---|---|
| Whole sample | 0.69 | 0.73 | 0.77 |
| Men | 0.70 | 0.76 | 0.77 |
| Women | 0.67 | 0.70 | 0.78 |
| Strong partisans | 0.93 | 0.95 | 0.95 |
| Weak partisans | 0.45 | 0.46 | 0.63 |
| Leaners | 0.70 | 0.74 | 0.79 |
| Conservatives | 0.59 | 0.66 | 0.71 |
| Moderates | 0.40 | 0.52 | 0.48 |
| Liberals | 0.43 | 0.25 | 0.51 |
| Whites | 0.64 | 0.70 | 0.78 |
| Blacks | 0.31 | 0.51 | 0.49 |
| Hispanics | 0.63 | 0.73 | 0.63 |
| Attends church | 0.71 | 0.75 | 0.77 |
| Doesn't attend church | 0.64 | 0.67 | 0.76 |
| Very interested in politics | 0.80 | 0.85 | 0.84 |
| Not at all interested in politics | 0.38 | 0.48 | 0.62 |
| Discusses politics | 0.72 | 0.76 | 0.79 |
| Doesn't discuss politics | 0.57 | 0.57 | 0.59 |
| 18 to 30 years old | 0.66 | 0.69 | 0.70 |
| 31 to 45 years old | 0.65 | 0.72 | 0.78 |
| 46 to 60 years old | 0.69 | 0.72 | 0.74 |
| Over 60 | 0.71 | 0.75 | 0.82 |

Cohen's Kappa findings show that there is a level of consensus, in areas indicating that GPT 3 can effectively imitate the responses of human evaluators. Individuals with affiliations and a keen interest in politics demonstrate the greatest level of agreement whereas individuals with weaker affiliations and moderates exhibit lower levels of agreement hinting at variations, in forecasting precision based on different levels of political involvement.

*Intraclass Correlation Coefficient (ICC).* This metric determines the consistency of ratings within groups to assess reliability. Comprehensive ICC values can be found in Table 4.

**Table 4: ICC values for various groups for the years 2012, 2016, and 2020 [2, Table 8, 9, 10]**

| Variable | 2012 | 2016 | 2020 |
|---|---|---|---|
| Whole sample | 0.81 | 0.84 | 0.87 |
| Men | 0.82 | 0.86 | 0.87 |
| Women | 0.80 | 0.82 | 0.88 |
| Strong partisans | 0.96 | 0.97 | 0.97 |
| Weak partisans | 0.61 | 0.62 | 0.77 |
| Leaners | 0.82 | 0.85 | 0.88 |
| Conservatives | 0.74 | 0.79 | 0.83 |
| Moderates | 0.57 | 0.69 | 0.65 |
| Liberals | 0.60 | 0.39 | 0.67 |
| Whites | 0.77 | 0.83 | 0.88 |
| Blacks | 0.47 | 0.67 | 0.66 |
| Hispanics | 0.78 | 0.85 | 0.77 |
| Attends church | 0.83 | 0.86 | 0.87 |
| Doesn't attend church | 0.78 | 0.80 | 0.86 |
| Very interested in politics | 0.89 | 0.92 | 0.91 |
| Not at all interested in politics | 0.53 | 0.64 | 0.77 |
| Discusses politics | 0.84 | 0.80 | 0.88 |
| Doesn't discuss politics | 0.73 | 0.72 | 0.74 |
| 18 to 30 years old | 0.80 | 0.81 | 0.82 |
| 31 to 45 years old | 0.79 | 0.84 | 0.88 |
| 46 to 60 years old | 0.82 | 0.83 | 0.85 |
| Over 60 | 0.83 | 0.85 | 0.90 |

The findings, from the ICC tests reveal agreement among groups suggesting that GPT 3s predictions are dependable within these specified categories. Individuals with leanings and a keen interest in politics exhibit the highest ICC values showcasing the models consistent performance for these subsets. Nevertheless lower values observed among moderates and specific demographic segments like individuals and those, with political engagement suggest areas where the models reliability could be enhanced.

*Exactly Match.* It signifies the ratio of agreement to ratings. The section discussing vote prediction and analysis of responses includes informative values (Table 5).

The findings, from the Proportion Agreement analysis suggest an uptick in alignment between GPT 3 forecasts and the ANES likelihood of supporting the presidential candidate climbing from 0.85 in 2012 to 0.89 in 2020 across the entire sample. Strong partisans consistently demonstrate the agreement maintaining a 0.97 reflecting a clear resonance with their voting tendencies. Both men and women show levels of concurrence with women edging higher at 0.90 by 2020. Weak partisans and individuals with an interest in politics also exhibit escalating levels of agreement. On the hand independents and moderates showcase lower and more fluctuating agreement rates, independents dropping to 0.53 by 2020. Within groups Blacks and liberals display notable agreement levels particularly liberals achieving a high of 0.97 in 2020. Whites and Hispanics also show alignment albeit with Hispanics experiencing a dip in accord for 2020 data points. Attendance at services does not notably impact agreement levels while older age brackets (over 60) present the level of concordance in predictions for the year 2020 (at 0.91) indicating consistent trends, across these age cohorts.

**Table 5: Exactly Match between GPT-3 and ANES probability of voting for the Republican presidential candidate [2, Table 1].**

| Variable | 2012 | 2016 | 2020 |
|---|---|---|---|
| Whole sample | 0.85 | 0.87 | 0.89 |
| Men | 0.85 | 0.88 | 0.88 |
| Women | 0.86 | 0.86 | 0.90 |
| Strong partisans | 0.97 | 0.97 | 0.97 |
| Weak partisans | 0.74 | 0.74 | 0.82 |
| Leaners | 0.85 | 0.87 | 0.89 |
| Independents | 0.59 | 0.62 | 0.53 |
| Conservatives | 0.84 | 0.86 | 0.89 |
| Moderates | 0.77 | 0.78 | 0.77 |
| Liberals | 0.95 | 0.95 | 0.97 |
| Whites | 0.82 | 0.85 | 0.89 |
| Blacks | 0.97 | 0.96 | 0.94 |
| Hispanics | 0.86 | 0.90 | 0.83 |
| Attends church | 0.86 | 0.88 | 0.88 |
| Doesn't attend church | 0.85 | 0.85 | 0.90 |
| High interest in politics | 0.90 | 0.93 | 0.92 |
| Low interest in politics | 0.74 | 0.75 | 0.81 |
| Discusses politics | 0.87 | 0.88 | 0.90 |
| Doesn't discuss politics | 0.82 | 0.79 | 0.79 |
| 18 to 30 years old | 0.87 | 0.86 | 0.87 |
| 31 to 45 years old | 0.85 | 0.87 | 0.90 |
| 46 to 60 years old | 0.86 | 0.86 | 0.87 |
| Over 60 | 0.85 | 0.87 | 0.91 |

*3.1.3 Bias and Ethical Considerations.* The research pointed out biases:

- **Demographic Bias**: Disparities, in model performance among demographic groups.
- **Ideological Bias**: Differences in bias towards distinct political ideologies.

Ethical considerations involve the risk of LLMs perpetuating existing biases and the importance of using these models responsibly in social science studies.

*3.1.4 Key Contributions.* The study showcased " fidelity " illustrating how GPT 3 can accurately mimic human sub groups by leveraging detailed socio demographic data. This indicates that LLMs can act as representations, for sub populations.

## 3.2 Santurkar et al. [15] - Whose Opinions Do Language Models Reflect?

Santurkar et al. [15] studied how well language models align, with the opinions of groups in the United States using the OpinionsQA dataset.

*3.2.1 Data and Methodology.* This dataset contains opinion questions and responses from a variety of US groups. They assessed how closely responses generated by language models like GPT 3 and GPT 4 matched the opinions of these groups [15, Figure 1].

*3.2.2 Performance Metrics.* The evaluation considered metrics:

*Wasserstein Distance:* This metric measured the difference between human and language model opinion distributions for questions utilizing the structured options in Pew surveys [15, Section A.4].

*Consistency Score (Cm):* It gauged how consistently a model reflects a particular groups opinions across different topics (Table 6).

$$C_m = \frac{1}{T} \sum_{T} \mathbf{1} \left( \arg\max_{G} R_M^G(Q_T) = G_{\text{best}}^m \right)$$

where:

- $T$ is the number of topics.
- $Q_T$ is the set of questions related to topic $T$.
- $R_M^G(Q_T)$ is the representativeness score of model $M$ for group $G$ on topic $T$.
- $G_{\text{best}}^m$ is the group for which the model $m$ has the highest overall alignment across all topics.
- $\mathbf{1}$ is the indicator function, which is 1 if the condition inside is true and 0 otherwise.

*Representativeness Score (RO):.* This metric examined how well a language models default opinion distribution aligns with that of the overall population or a specific demographic group [15, Figure 2].

$$R_m^O(Q) = A(D_m, D_O, Q)$$

where:

- $D_m$ is the opinion distribution of the model.
- $D_O$ is the opinion distribution of the overall population or a specific demographic group.
- $Q$ is the set of questions considered.
- $A$ is the alignment measure, calculated as $A(D_1, D_2) = 1 - \frac{\text{WD}(D_1, D_2)}{N-1}$, where WD is the Wasserstein distance between distributions $D_1$ and $D_2$, and $N$ is the number of answer choices.

**Table 6: Performance metrics for different language models [15, Figure 6, Figure 2].**

| Model | Consistency Score (Cm) | Representativeness Score (RO) |
|---|---|---|
| j1-grande | 0.612 | 0.813 |
| j1-jumbo | 0.612 | 0.816 |
| j1-grande-v2-beta | 0.575 | 0.804 |
| ada | 0.622 | 0.824 |
| davinci | 0.562 | 0.791 |
| test-ada-001 | 0.388 | 0.707 |
| text-davinci-001 | 0.388 | 0.707 |
| text-davinci-002 | 0.502 | 0.763 |
| text-davinci-003 | 0.575 | 0.700 |

The data, in Table 6 comparing language models reveals variations in how well they capture group sentiments consistently and match overall population trends. Models like j1 grande and j1 jumbo stand out with Consistency Scores (Cm) of 0.612 indicating their

reliability in reflecting group opinions across subjects. On the hand text davinci 002 exhibits the Consistency Score of 0.502 suggesting it struggles to maintain consistency in representing group viewpoints. In terms of Representativeness Score (RO) which gauges how a models default opinion distribution aligns with the survey population, j1 jumbo and j1 grande lead again with scores of 0.816 and 0.813 respectively showing their alignment with overall population preferences. Conversely text davinci 003 and text davinci 002 score lower on RO (0.700 and 0.763) indicating a fit with the survey populations views. These findings imply that for tasks demanding both representation and alignment, with opinions j1 grande and j1 jumbo are the more suitable choices.

*3.2.3 Bias and Ethical Considerations.* The study also addressed biases and ethical concerns:

- **Demographic Bias**: Some groups, like individuals aged 65+ and widowed individuals were found to be underrepresented.
- **Ideological Bias**: LLMs often show left leaning tendencies due to biases, in their training data.

The research highlights the importance of enhancing alignment methods and conducting bias evaluations to guarantee representation of various demographic viewpoints, in LLM results.

*3.2.4 Key Contributions.* In their work Santurkar et al. [15] introduced OpinionsQA to assess how well LLMs align with the viewpoints of 60 groups in the US. Their findings unveiled discrepancies and enduring biases despite efforts to direct models, towards demographics.

## 3.3 Durmus et al. [5] - Towards Measuring the Representation of Subjective Global Opinions in Language Models

Durmus et al. [5] examined how well Language Models (LLMs) represent opinions using the GlobalOpinionQA dataset to understand diverse perspectives on societal issues worldwide.

*3.3.1 Data and Methodology.* The GlobalOpinionQA dataset comprises 2,556 multiple choice questions sourced from the Pew Research Centers Global Attitudes surveys and the World Values Survey. The study involved presenting these questions to an LLM and comparing the distribution of model generated responses, with responses across countries using the Jensen Shannon Distance [5, Figure 1].

*3.3.2 Performance Metrics.* Model performance was measured by the similarity between model-generated and human responses using the following metrics:

*Jensen-Shannon Distance.* The performance of the model was assessed based on the similarity between its generated responses and human answers using key metrics; Jensen Shannon Distance; This metric measures how closely aligned the probability distributions of model generated responses are with those of humans. The outcomes are detailed in Table 7.

**Table 7: Performance Metrics for Model Response Similarity (extracted from https://llmglobalvalues.anthropic.com.**

| Metric | USA | Russia | China | Australia | Pakistan |
|---|---|---|---|---|---|
| Default Prompting | 0.68 | 0.66 | 0.61 | 0.67 | 0.61 |
| Cross National Prompting (Germany) | 0.62 | 0.60 | 0.60 | 0.61 | 0.56 |
| Cross National Prompting (China) | 0.66 | 0.70 | 0.70 | 0.64 | 0.68 |
| Cross National Prompting (Russia) | 0.68 | 0.72 | 0.67 | 0.66 | 0.68 |
| Linguistic Prompting (Russian) | 0.67 | 0.66 | 0.61 | 0.66 | 0.62 |
| Linguistic Prompting (Chinese) | 0.67 | 0.68 | 0.64 | 0.68 | 0.62 |

The results, from Jensen Shannon Distance offer insights into how model generated responses match answers in different countries. The Default Prompting method generally shows alignment with distances ranging between 0.61 and 0.68 indicating a not perfect fit. When using Cross National Prompting focusing on Germany leads to improved alignment for countries for Pakistan (0.56). However selecting China and Russia for Cross National Prompting boosts alignment significantly for those countries reaching up to 0.70 and 0.72 respectively. Linguistic Prompting follows a pattern with Russian and Chinese prompting resulting in alignment for their respective languages and slight improvements for other nations. Overall strategies like Cross National Prompting and Linguistic Prompting help enhance the models alignment, with patterns particularly when the prompts closely match the language or cultural context of the target country.

*3.3.3 Bias and Ethical Considerations.* The research uncovered biases in LLM responses that tended to favor viewpoints, from Industrialized, Rich and Democratic (WEIRD) populations [5, Figure 2]. Ethical concerns encompass the risk of perpetuating stereotypes and the significance of incorporating a range of perspectives, in model creation.

*3.3.4 Key Contributions.* The study presented GlobalOpinionQA, a system designed to assess how well LLMs align, with viewpoints uncovering biases favoring cultures and underscoring obstacles in integrating non Western outlooks.

## 3.4 Bisbee et al. [3] - Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

Bisbee et al. [3] examined how LLMs, ChatGPT were used to create survey data with a focus, on the challenges and limitations encountered.

*3.4.1 Data and Methodology.* The research utilized ChatGPT 3.5 Turbo, which was programmed to take on personas based on political traits. The synthetic responses were compared against survey data from the American National Election Study (ANES) spanning from 2016 to 2020 [3, Figure 2]. The study concentrated on three areas; capturing the sentiments towards various groups exploring the connections between persona characteristics and responses and assessing how changes in prompts and model updates impacted the outcomes.

*3.4.2 Performance Metrics.* The performance assessment involved measuring Mean Absolute Error (MAE) Standard Deviation and F1 Score [3, Figure 2].

**Table 8: F1 Scores for various government role questions [3, Figure 60]**

| Question | F1 Score |
|---|---|
| Government financing of projects to create new jobs | 0.80 |
| Support for industry to develop new products and technology | 0.75 |
| Support for declining industries to protect jobs | 0.71 |
| Cuts in government spending | 0.51 |
| Less government regulation of business | 0.48 |
| Reducing the working week to create more jobs | 0.51 |

The data, on performance measures show that different government role queries exhibit varying levels of effectiveness. The top F1 Score of 0.80 is achieved for the query regarding government funding for projects aimed at generating employment opportunities indicating the models predictive capability in this domain. Similarly providing assistance to industries for developing products and technology well as aiding declining sectors to safeguard jobs display relatively high F1 Scores of 0.75 and 0.71 respectively signifying commendable performance. On the hand inquiries concerning reductions in government expenditures decreased control over businesses and shortening the workweek to enhance job creation yield lower F1 Scores ranging from 0.48 to 0.51 implying weaker predictive accuracy in these areas. These findings underscore that while the model excels in forecasting support for government functions its efficacy diminishes when addressing scenarios involving a decrease, in government involvement.

*3.4.3 Replicability and Bias.* Several biases and reproducibility concerns were identified in the study:

- **Overconfidence in Responses**: Synthetic data exhibited variability compared to survey data resulting in an overestimation of statistical conclusions.
- **Sensitivity to Prompts**: Even slight adjustments in prompt wording or timing had a significant impact on the distribution of synthetic responses.
- **Reproducibility Concerns**: Findings varied following updates to the LLM sparking concerns, about replicating results over time.

*3.4.4 Key Contributions.* Bisbee et al. [3] conducted an analysis of the advantages and drawbacks of employing LLMs for producing survey data. They emphasized the hurdles in maintaining the excellence, dependability and replicability of data advising against its substitution, for conventional survey techniques, in social science studies.

## 3.5 Kim and Lee [9] - AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys

In Kim and Lee [9] the application of language models (ALMs) to enhance surveys for opinion prediction was explored using data from the General Social Survey (GSS) spanning from 1972 to 2021 with a focus on retrodiction and predicting opinions.

*3.5.1 Data and Methodology.* The research made use of the GSS dataset containing 3,110 opinions from 68,846 individuals. A tuned Alpaca 7B model [11] was utilized to forecast survey responses by incorporating question meanings, individual beliefs and temporal contexts [9, Figure 2].

*3.5.2 Performance Metrics.* The performance of the model was assessed using metrics such as Area Under Curve (AUC) accuracy and F1 score for tasks like imputing missing data, retrodiction and predicting opinions (Table 9)[9, Figure 3].

**Table 9: Comparison of prediction performances across four different models across three scenarios [9, Table A4].**

| Models | Alpaca-7b | GPT-J-6b | RoBERTa-large | Matrix Factorization |
|---|---|---|---|---|
| Missing data imputation: | | | | |
| - AUC | 0.866 | 0.864 | 0.859 | 0.852 |
| - Accuracy | 0.782 | 0.779 | 0.774 | 0.784 |
| - F1-score | 0.765 | 0.765 | 0.758 | 0.770 |
| Retrodiction: | | | | |
| - AUC | 0.860 | 0.859 | 0.853 | 0.798 |
| - Accuracy | 0.775 | 0.774 | 0.768 | 0.740 |
| - F1-score | 0.755 | 0.759 | 0.750 | 0.687 |
| Unasked opinion prediction: | | | | |
| - AUC | 0.729 | 0.687 | 0.566 | |
| - Accuracy | 0.667 | 0.632 | 0.546 | |
| - F1-score | 0.640 | 0.609 | 0.422 | |

The results, from Table 9 show how different models perform across tasks like filling in missing data predicting events and anticipating opinions. In the task of imputing missing data Alpaca 7b stands out with the AUC (0.866) and accuracy (0.782) indicating its effectiveness. On the hand Matrix Factorization slightly edges ahead in terms of F1 score (0.770). When it comes to predicting events Alpaca 7b leads with an AUC of 0.860 demonstrating predictive abilities, closely followed by GPT J 6b. In predicting opinions Alpaca 7b maintains the spot with an AUC of 0.729 and an F1 score of 0.640; however its overall accuracy and F1 scores are lower compared to other tasks due to the complexity involved in this type of prediction. These findings highlight the strengths of Alpaca 7b and GPT J

6b, in tasks while also pointing out areas that could be enhanced especially when it comes to predicting unspoken opinions.

*3.5.3 Bias and Ethical Considerations.* The study highlighted biases in LLMs predictions:

- **Demographic Bias**: Predictions exhibit variations among different demographic groups potentially introducing biases.
- **Temporal Bias**: Changes in word meanings and cultural contexts over time may impact prediction accuracy, particularly with historical data.

Ethical considerations encompass addressing reinforcement of existing biases and safeguarding privacy well as individual autonomy, during opinion prediction.

*3.5.4 Key Contributions.* The research successfully attained a level of precision, in predicting events and unexpressed viewpoints by refining large language models through multiple cross sectional surveys.

## 3.6 Chu et al. [4] - Language Models Trained on Media Diets

Chu et al. [4] examined the use of LLMs trained on media consumption patterns to forecast sentiment. They adjusted these models to mimic opinions of groups, within the population based on their media preferences.

*3.6.1 Data and Methodology.* The research dataset consisted of surveys representative of the population regarding COVID 19 and consumer confidence. The approach involved customizing LLMs to align with media consumption patterns assessing their capacity to anticipate survey responses linked to media usage [4, Figure 1].

*3.6.2 Performance Metrics.* The performance assessment utilized correlation and regression analyses.

*Correlation:* The correlation between media consumption scores and survey response ratios was determined as $r = 0.458$. Regarding consumer confidence queries predictions following a approach displayed a correlation of $r = 0.376$ while sociocentric prospective predictions showed $r = 0.264$ [4, Figure 2].

*Regression:* Using media consumption scores regression analysis predicted survey response ratios yielding a regression coefficient $\beta = 0.115$. The application of a linear general additive model enhanced prediction accuracy (error=0.161) compared to the baseline BERT model (error=0.173) [4, Figure 2].

**Table 10: Performance Metrics for Media Diet Models**

| Domain | Correlation |
|---|---|
| COVID-19 | 0.458 |
| Consumer Confidence (Sociocentric-Retrospective) | 0.376 |
| Consumer Confidence (Sociocentric-Prospective) | 0.264 |

The results, from the performance metrics show how media consumption scores can predict survey responses. There is a connection ($r = 0.458$) between media consumption and COVID 19 response rates indicating that peoples media habits play a role in shaping their views on this issue. When it comes to predicting consumer confidence there are correlations. $r = 0.376$ for sociocentric retrospective and $r = 0.264$ for sociocentric suggesting decent predictive accuracy but with room for improvement. Through regression analysis using media consumption scores we find a regression coefficient of $\beta = 0.115$ indicating a link between media habits and survey responses. By applying a additive model we enhance prediction accuracy reducing errors from 0.173 to 0.161 and showing better performance compared to the baseline BERT model. These findings emphasize the significance of including media consumption patterns in models for survey responses for topics heavily influenced by media coverage, like COVID 19.

*3.6.3 Bias and Ethical Considerations.* Chu et al. [4] found certain biases and ethical considerations to take into account:

- **Media Bias**: The study pointed out that LLMs might exhibit biases found in the media material they are trained on.
- **Ethical Implications**: The use of media diet models raises dilemmas concerning perpetuating media biases and shaping viewpoints.

*3.6.4 Key Contributions.* Chu et al. [4] introduced the concept of media diet models showcasing that LLMs tailored to media content can anticipate opinion based on media consumption, with accuracy and resilience to variations in media exposure. The research emphasized how the media influences opinion and underscored the necessity for exploration, into the ethical ramifications of employing LLMs for opinion prediction.

## 3.7 Hwang et al. [7] - Aligning Language Models to User Opinions

Hwang et al. [7] delved into how language models (LLMs) align, with user opinions using the OpinionQA dataset. Their research aimed to improve prediction accuracy by considering factors like age, gender, race, education and ideological viewpoints such as social beliefs.

*3.7.1 Data and Methodology.* The team trained LLMs on the OpinionQA dataset to predict user opinions. They utilized sets of demographic and opinion data combinations to assess model accuracy [7, Figure 1]).

*3.7.2 Performance Metrics.* They evaluated performance using metrics like match accuracy collapsed match accuracy and Cohens kappa coefficient. Exact match accuracy gauges how many predictions precisely match user responses while collapsed match accuracy combines answer selections (Table 11). Cohens kappa coefficient quantified agreement levels among users with demographics but differing opinions [7, Figure 2].

The performance metrics show how different model setups predict user reactions. The model setup combining Demographic, Ideology and Top 8 Opinions achieves the exact match accuracy at 0.54 indicating that including these elements boosts precision. This configuration also has the collapsed match accuracy of 0.70 implying that grouping answer choices improves performance further. Other setups, like Demographic + Ideology and Demographic + Top 8 Opinions also perform well with match accuracies of 0.53 and collapsed match accuracies of 0.69 respectively. In contrast the

**Table 11: Exact and Collapsed Match Accuracy for various model configurations [7, Table 3].**

| Model | Exact Match Accuracy | Collapsed Match Accuracy |
|---|---|---|
| No Persona | 0.43 | 0.62 |
| Demographic + Ideology | 0.47 | 0.65 |
| Demographic + Ideology + All Opinions | 0.51 | 0.69 |
| Ideology + Top-8 Opinions | 0.53 | 0.69 |
| Demographic + Top-8 Opinions | 0.53 | 0.69 |
| Demographic + Ideology + Top-3 Opinions | 0.53 | 0.69 |
| Top-3 Opinions | 0.51 | 0.67 |
| Top-8 Opinions | 0.52 | 0.68 |
| **Demographic + Ideology + Top-8 Opinions** | **0.54** | **0.70** |

baseline model without any persona data has the match accuracy at 0.43 and a collapsed match accuracy of 0.62 underscoring the significance of incorporating demographic and opinion information. Cohens kappa scores hovering around 0.4 suggest a level of consensus among users with demographics but differing opinions indicating some correlation but also variability in individual responses. These findings underscore the importance of opinion data in improving model accuracy and hint at enhancements, by fine tuning these factors.

*3.7.3 Bias and Ethical Considerations.* The study highlighted biases in LLM predictions:

- **Demographic Bias**: Models exhibit varying performance across segments with certain groups receiving more accurate forecasts than others.
- **Ideological Bias**: Incorporating ideological details can enhance prediction precision. Might also perpetuate existing biases.

Ethical considerations encompassed the risk of LLMs reinforcing stereotypes. Emphasized the need for fairness, in model predictions.

*3.7.4 Key Contributions.* The research showed that by incorporating information and previous opinions the accuracy of predicting user opinions, in LLMs was significantly enhanced.

## 3.8 Sanders et al. [14] - Demonstrations of the Potential of AI-based Political Opinion Analysis

In their study Sanders et al. [14] delved into the use of language models with a focus, on GPT 3.5 for analyzing opinions through simulated survey responses that mirror public sentiments on diverse political matters.

*3.8.1 Data and Methodology.* The researchers utilized a dataset comprising polling questions on policy issues from the 2022 Cooperative Election Study (CES) a survey involving around 60,000 US respondents chosen to represent the nation. By feeding ideological inputs GPT 3.5 generated responses for comparison with data collected through the CES.

*3.8.2 Performance Metrics.* To assess how well the language model replicated public opinion responses, performance metrics like the Pearson correlation coefficient ($\rho$). Mean Absolute Percentage Error (MAPE) were employed.

**Table 12: Performance Metrics for Different Issues [14, Figure 6]**

| Issue | Demographic Fields | Pearson Correlation ($\rho$) | MAPE (%) |
|---|---|---|---|
| SCOTUS Approval | All | 92.1% | 9.3% |
| SCOTUS Approval | Ideology | 96.6% | 7.3% |
| Police Safety | Ideology | 92.7% | 37.6% |
| Abortion Ban | Ideology | 98.3% | 6.4% |
| Increase Fuel Production | Ideology | 85.6% | 22.7% |
| Prescription Import | Ideology | 66.8% | 31.3% |
| Gun Background Checks | Ideology | 91.3% | 28.8% |

Pearson correlation ($\rho$) measures the linear association between AI generated and human survey answers. For instance approval ratings of SCOTUS across demographics displayed a 92.1% correlation with a MAPE of 9.3% improving to 96.6% correlation and 7.3% MAPE within ideological subsets. On topics like police safety there was a correlation of 92.7%. With a higher MAPE of 37.6% indicating discrepancies in absolute values (Table 12).

The performance metrics, for topics demonstrate varying levels of accuracy in reflecting public opinion responses. The Pearson correlation coefficients show linear connections within specific ideological groups. For instance approval ratings for SCOTUS show a correlation of 96.6% within subsets compared to 92.1% across all demographics indicating that the models predictions are more precise when considering ideological differences. The MAPE values, which assess prediction errors are relatively low for SCOTUS approval (7.3% within ideology) suggesting precision. However issues like police safety and prescription import exhibit higher MAPE values (37.6% and 31.3% respectively) pointing out disparities in the predicted values by the model. The correlation and MAPE results for abortion ban (98.3% and 6.4%) reveal performance. These findings highlight the models effectiveness, in forecasting topics with accuracy particularly when demographic and ideological aspects are taken into account while also pinpointing areas where prediction precision can be enhanced.

*3.8.3 Comparison with Human Data.* The AI generated outputs demonstrated correlations with data within ideological categories often surpassing 85% [14, Figure 2]. Nevertheless accuracy levels varied concerning variables such, as age, race and gender.

*3.8.4 Bias and Ethical Considerations.* The research highlighted biases observed in responses generated by AI:

- **Demographic Bias**: Inaccuracies are more prevalent, in predicting responses for specific demographic groups especially among non White individuals.

- **Ideological Bias**: There is a tendency to oversimplify ideological differences particularly when addressing new policy issues not included in the training data.

Ethical considerations encompass the risk of language models reinforcing existing biases and the necessity for transparency and accountability when utilizing these models for analyzing opinions.

*3.8.5 Key Contributions.* The research demonstrated that LLMs can effectively predict sentiment on topics particularly within ideological divisions although their effectiveness diminishes when addressing differences, at the demographic level.

## 3.9 Lee et al. [10] - Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias

Lee et al. [10] studied how well language models (LLMs) can predict sentiment regarding warming. They examined LLM performance by analyzing survey data from groups [10, Figure 6].

*3.9.1 Data and Methodology.* The dataset contained survey responses related to warming. The approach involved refining LLMs to forecast survey responses based on psychological factors.

*3.9.2 Performance Metrics.* The study measured performance using accuracy, F1 score and Macro Average F1 score (MAF1) (Tables 13 and 14).

**Table 13: Evaluation Metrics on Global Warming Belief with Binary Answer Option (Supplemental Table 4)**

| Model | Year | Accuracy | MAF1 Score | Precision | Recall |
|---|---|---|---|---|---|
| GPT 4.0 | 2017 | 0.89 | 0.82 | 0.79 | 0.86 |
| GPT 4.0 | 2021 | 0.91 | 0.85 | 0.82 | 0.92 |
| GPT 3.5 | 2017 | 0.80 | 0.53 | 0.55 | 0.53 |
| GPT 3.5 | 2021 | 0.86 | 0.65 | 0.76 | 0.62 |

**Table 14: Evaluation Metrics on Global Warming Belief with Multiple Answer Options (Supplemental Table 5)**

| Model | Year | Accuracy | MAF1 Score | Precision | Recall |
|---|---|---|---|---|---|
| GPT 4.0 | 2017 | 0.77 | 0.54 | 0.64 | 0.53 |
| GPT 4.0 | 2021 | 0.82 | 0.57 | 0.66 | 0.54 |
| GPT 3.5 | 2017 | 0.72 | 0.47 | 0.69 | 0.45 |
| GPT 3.5 | 2021 | 0.75 | 0.47 | 0.69 | 0.45 |

F1 and MAF1 scores were computed for subgroups as presented in Table 15. Notably GPT 4 conditioned on demographics and covariates exhibited MAF1 enhancements across years and survey scenarios.

The performance data suggests that GPT 4.0 performs better, than GPT 3.5 in both multiple choice answer situations. In terms of answers (see Table 13) GPT 4.0 achieves accuracy and MAF1 scores with the 2021 results showing a 0.91 accuracy and 0.85 MAF1 indicating improved predictive abilities over time. Despite improvements from 2017 to 2021 GPT 3.5 falls short compared to GPT 4.0 across all metrics.

In scenarios with answer choices (refer to Table 14) GPT 4.0 also exhibits performance albeit with slightly lower overall accuracy and MAF1 scores compared to the binary options reflecting the tasks increased complexity. Noteworthy is the improvement in accuracy and MAF1 scores for GPT 4.0 from 2017 to 2021 suggesting model enhancements over time.

The calculated F1 scores based on race/ethnicity (see Table 15) demonstrate performance by GPT4.0 among non Hispanic Whites and those identifying with two or more races non Hispanic individuals. However Blacks non Hispanic individuals show scores particularly for negative responses pointing towards a potential area, for enhancement. In general the data shows that GPT 4.0 has improved in predicting sentiment among groups of people but there is still room, for improvement to better address differences, within certain subgroups.

*3.9.3 Bias and Ethical Considerations.* The research presented specific biases:

- **Demographic Bias**: The model downplays concerns, within certain demographics like Black Americans.
- **Temporal Bias**: Evolving word meanings and cultural contexts over time can impact prediction precision.

Ethical considerations encompass the risk of LLMs perpetuating existing biases as the necessity of ensuring fairness and accuracy in model predictions.

*3.9.4 Key Contributions.* Lee et al. [10] evaluated how accurately LLMs capture sentiment on warming while highlighting aspects, like model choice, prompt design and bias evaluation.

## 3.10 Sun et al. [17] - Random Silicon Sampling: Simulating Human Sub-Populations

Sun et al. [17] investigated the application of Random Silicon Sampling (RSS) to mimic sub groups assessing how well Language Learning Models (LLMs) can create samples that represent a variety of sub population characteristics.

*3.10.1 Data and Methodology.* As presented in Sun et al. [17, Figure 1], the research utilized opinion data sources to generate synthetic datasets through RSS mirroring real world population distributions. The models were assessed based on their capability to mirror distributions and opinion patterns.

*3.10.2 Performance Metrics.* The effectiveness of the RSS approach was measured using the Chi Square Test, for Homogeneity and Kullback Leibler (KL) Divergence.

*Chi-Square Test for Homogeneity.* This assessment compared response distributions between ANES data and generated samples highlighting disparities ($p < 0.05$). Results are shown in Table 16.

**Table 15: Synthesized F1 Scores of GPTs for Binary Belief in Global Warming by Race/Ethnicity**

| Race/Ethnicity | Models | Year | F1 Score (Yes) | F1 Score (No) |
|---|---|---|---|---|
| 2+ Races, Non-Hispanic | GPT 4.0 | 2017 | 0.96 | 0.75 |
| 2+ Races, Non-Hispanic | GPT 4.0 | 2021 | 1 | 1 |
| Black, Non-Hispanic | GPT 4.0 | 2017 | 0.95 | 0.29 |
| Black, Non-Hispanic | GPT 4.0 | 2021 | 0.96 | 0.25 |
| Hispanic 2017 | GPT 4.0 | 2017 | 0.94 | 0.61 |
| Hispanic 2017 | GPT 4.0 | 2021 | 0.95 | 0.96 |
| Other, Non-Hispanic | GPT 4.0 | 2017 | 0.98 | 0.83 |
| Other, Non-Hispanic | GPT 4.0 | 2021 | 0.94 | 0.33 |
| White, Non-Hispanic | GPT 4.0 | 2017 | 0.92 | 0.72 |
| White, Non-Hispanic | GPT 4.0 | 2021 | 0.94 | 0.78 |

**Table 16: Replicability of Random Silicon Sampling [17, Table 1]**

| Sample | Biden Rate | Trump Rate | $\chi^2$ |
|---|---|---|---|
| ANES 2020 | 58.88% | 41.18% | - |
| Silicon Sample | 55.61% | 44.39% | 8.8931 |
| RSS 1 | 58.00% | 42.00% | 0.5688 |
| RSS 2 | 57.99% | 42.01% | 0.5897 |
| RSS 3 | 57.85% | 42.15% | 0.8107 |
| RSS 4 | 57.73% | 42.27% | 1.0182 |
| RSS 5 | 57.45% | 42.55% | 1.5668 |
| RSS 6 | 57.27% | 42.73% | 2.0724 |
| RSS 7 | 57.24% | 42.76% | 2.1671 |
| RSS 8 | 57.19% | 42.81% | 2.3121 |
| RSS 9 | 57.02% | 42.98% | 2.8054 |
| RSS 10 | 56.61% | 43.39% | 4.2774 |

The outcomes of the Chi Square Test, for Homogeneity reveal how closely the response distributions in samples mirror those in the ANES 2020 dataset. The Silicon Sample stands out with a divergence indicated by a $\chi^2$ value of 8.8931 implying a difference from the ANES data. In contrast the Random Silicon Samples (RSS) display lower $\chi^2$ values ranging from 0.5688 to 4.2774 across RSS 1 to RSS 10. This suggests a resemblance to the ANES distribution evident in RSS 1 and RSS 2 with minimal disparities ($\chi^2$ values of 0.5688 and 0.5897 respectively).

As we observe increases in $\chi^2$ values across the RSS samples the distinctions between these samples and the ANES data become more noticeable. Still remain relatively modest compared to the Silicon Samples disparity. These findings imply that employing silicon sampling techniques can result in distributions that closely mirror those of ANES data thereby improving the reproducibility of public opinion responses generated by the model. Overall the lower $\chi^2$ values in RSS samples underscore their efficacy in emulating survey data affirming the trustworthiness of these sampling methods, for producing datasets.

*Kullback-Leibler Divergence.* KL Divergence values quantified the resemblance between ANES data and RSS response distributions, where lower values indicated similarity [17, Figure 2].

**Table 17: KL-Divergence of Random Silicon Sampling [17, Table 1]**

| Sample | KL-Divergence |
|---|---|
| Silicon sample | 0.00210 |
| RSS 1 | 0.00014 |
| RSS 2 | 0.00014 |
| RSS 3 | 0.00020 |
| RSS 4 | 0.00024 |
| RSS 5 | 0.00039 |
| RSS 6 | 0.00049 |
| RSS 7 | 0.00051 |
| RSS 8 | 0.00055 |
| RSS 9 | 0.00066 |
| RSS 10 | 0.00100 |

These metrics showcased how well the RSS method could replicate response distributions akin to those seen in the ANES dataset (Table 17).

The KL Divergence values offer insights, into how the response distributions from samples mirror the ANES data. Lower KL Divergence values indicate a resemblance. The Silicon Sample shows a KL Divergence of 0.00210 indicating some deviation from the ANES data. In contrast the Random Silicon Samples (RSS) display KL Divergence values, ranging from 0.00014 to 0.00100. RSS 1 and RSS 2 have the values (0.00014) suggesting the match to the ANES distributions.

As the KL Divergence values slightly increase across the RSS samples, their alignment with the ANES data becomes less perfect. Remains significantly better than that of the Silicon Sample. These findings illustrate that the RSS approach can effectively replicate response distributions to those seen in the ANES dataset. The ability of RSS to generate KL Divergence values underscores its reliability in producing data validating its effectiveness, in public opinion research.

*3.10.3 Replicability and Bias.* The study pinpointed biases affecting reproducibility:

- **Harmlessness Bias**: The model tends to offer non-controversial answers on sensitive topics.
- **Demographic Bias**: It exhibited higher accuracy, in replicating younger and more educated demographics compared to older and less educated groups.

Ethical concerns involve the possibility of LLMs misrepresenting communities and the importance of validating artificial data.

*3.10.4 Key Contributions.* In their study, Sun et al. [17] introduced the RSS approach, for simulating subgroups based on information showcasing the practicality as well as biases present, in LLMs.

# 4 Discussion

Performance results are presented in Table 18 which give us a summary of how well different studies have assessed the effectiveness of Language Models (LLMs), in various situations. Each metric highlights both the strengths and weaknesses of these models in capturing feedback and aligning with real world data.

For example metrics such as Tetrachoric Correlation, Cohen's Kappa, Intraclass Correlation Coefficient (ICC) and Proportion Agreement as discussed by Argyle et al. [2] show a level of accuracy in replicating survey responses using LLMs. In some cases scores go up to 1.00 indicating agreement. These metrics help evaluate how effectively LLMs can imitate decision making processes and account for differences among socio demographic groups.

Furthermore metrics like Consistency Score (Cm) and Representativeness Score (RO) from Santurkar et al. [15] highlight the ability of LLMs to maintain responses across demographic categories. The Jensen Shannon Distance metric, as explained by Durmus et al. [5] measures how model generated responses resemble input providing insights, into how well these models can replicate human perspectives.

In addition the metrics, like Correlation and F1 Score studied by Chu et al. [4] and Bisbee et al. [3] respectively show how well models can predict outcomes in situations, such as trends in media consumption and creating survey data. The importance of metrics like Area Under Curve (AUC) and Accuracy emphasized by Kim and Lee [9] is in evaluating how effectively the models deal with missing data and forecast opinions showcasing their robustness in handling datasets.

Furthermore, the metrics of Pearson Correlation and MAPE from Sanders et al. [14] along with Kullback Leibler Divergence (KL) and Chi Square Test for Homogeneity from Sun et al. [17] offer insights into how the models correlate with sentiments and responses specific to various demographics. These metrics stress the necessity, for refining and assessing to reduce biases and enhance the reliability of data produced by LLMs.

The analysis depicted in Table 18 highlights the applications and effectiveness of LLMs across fields. By delving into a range of metrics this research provides perspectives on the capabilities of these models while also identifying areas that could benefit from improvement. As LLMs advance, continuous evaluation and adjustments will be essential to enhance their accuracy reduce biases and maintain their credibility, in social science research.

## 4.1 Summary of Key Points

Upon examining the papers it is clear that there are viewpoints on the efficacy of LLMs in survey studies. The findings from these papers typically categorize into two groups:

- Papers indicating that LLMs are effective ("they work").
- Papers highlighting limitations or challenges ("they don't work").

*4.1.1 Papers Supporting Effectiveness.* Numerous studies offer proof to validate the notion that LLMs can effectively model opinions:

- Hwang et al. [7] demonstrate enhanced accuracy in predicting opinions by integrating data and past opinions yielding an Exact Match Accuracy of 0.54 and a Collapsed Match Accuracy of 0.70.
- Kim and Lee [9] achieve precision in retrodiction and opinion prediction through tuning LLMs with repeated sectional surveys attaining an AUC of 0.866 and an accuracy rate of 0.782.
- Durmus et al. [5] showcase that LLMs can capture perspectives using the GlobalOpinionQA dataset achieving a Jensen Shannon Distance of 0.56.
- Sanders et al. [14] highlight that AI chatbots powered by language models can accurately predict sentiments, on topics from different perspectives with a high correlation coefficient of 0.983 and a low MAPE value of 0.376.
- Lee et al. [10] delved deeper into how these models can forecast opinion on warming achieving an accuracy rate of 0.91 and an F1 Score of 0.765.

*4.1.2 Papers Highlighting Limitations and Challenges.* Various research papers have also highlighted the challenges and limitations associated with using Language Models in opinion surveys:

- Argyle et al. [2] discussed biases in representations related to race and gender within Language Models indicating that while they can replicate groups biases still exist, as seen from Tetrachoric Correlation values ranging between 0.65 to 1.00 and Cohen's Kappa values varying from 0.25 to 0.95.
- Sun et al. [17] pointed out issues of partiality in responses neutrality towards topics using the "silicon sampling" method revealing Kullback Leibler Divergence values as minimal as 0.00014.
- Chu et al. [4] showcased that although models analyzing media consumption patterns can predict opinions based on individuals content consumption habits biases are present, in how information's presented with correlation ranges between 0.264 to 0.458.
- Santurkar et al. [15] discovered discrepancies in perspectives among Language Models and different demographic segments with Consistency Scores ranging from 0.388 to 0.622.
- Bisbee et al. [3] explored the intricacies of utilizing data than real survey answers highlighting potential prejudices in forecasting demographics with F1 Scores fluctuating, between 0.48 and 0.80.

## 4.2 Synthesis of Biases Identified by Researchers

Researchers have identified biases in LLMs and discussed the ethical implications of using these models for survey research (Table 19).

*4.2.1 Details on Bias Metrics.* Different studies have used metrics to measure biases:

- Argyle et al. [2] used Cohens Kappa and Accuracy to assess the agreement between LLM predictions and real responses shedding light on biases related to race and gender.

| Metric | Paper | Best Score | Range | Ref |
|---|---|---|---|---|
| Tetrachoric Correlation | [2] | 1.00 | 0.65 - 1.00 | Table 2 |
| Cohen's Kappa | [2] | 0.95 | 0.25 - 0.95 | Table 3 |
| Intraclass Correlation Coefficient (ICC) | [2] | 0.97 | 0.39 - 0.97 | Table 4 |
| Exactly Match | [2] | 0.97 | 0.53 - 0.97 | Table 5 |
| Consistency Score (Cm) | [15] | 0.622 | 0.388 - 0.622 | Table 6 |
| Representativeness Score (RO) | [15] | 0.824 | 0.700 - 0.824 | Table 6 |
| Jensen-Shannon Distance | [5] | 0.56 | 0.56 - 0.72 | Table 7 |
| Correlation | [4] | 0.458 | 0.264 - 0.458 | Table 10 |
| F1 Score | [3] | 0.80 | 0.48 - 0.80 | Table 8 |
| Area Under Curve (AUC) | [9] | 0.866 | 0.566 - 0.866 | Table 9 |
| Accuracy | [9] | 0.782 | 0.546 - 0.782 | Table 9 |
| F1 Score | [9] | 0.765 | 0.422 - 0.765 | Table 9 |
| Pearson Correlation ($\rho$) | [14] | 0.983 | 0.668 - 0.983 | Table 12 |
| MAPE | [14] | 0.376 | 0.064 - 0.376 | Table 12 |
| Kullback-Leibler Divergence (KL) | [17] | 0.00014 | 0.00014 - 0.00210 | Table 17 |
| Chi-Square Test for Homogeneity ($\chi^2$) | [17] | 0.5688 | 0.5688 - 8.8931 | Table 16 |
| Exact Match Accuracy | [7] | 0.54 | 0.43 - 0.54 | Table 11 |
| Collapsed Match Accuracy | [7] | 0.70 | 0.62 - 0.70 | Table 11 |
| Cohen's Kappa | [7] | 0.57 | 0.42 - 0.57 | Hwang et al. [7, Figure 1] |

**Table 18: Consolidated Performance Metrics with Best Scores and Range of Results**

**Table 19: Summary of Biases Identified in LLM Research**

| Study | Type of Bias | Description |
|---|---|---|
| Argyle et al. [2] | Racial, Gender Bias | Biases in representation of racial and gender groups |
| Santurkar et al. [15] | Demographic Misalignment | Misalignment between LLM-generated and actual demographic opinions |
| Durmus et al. [5] | Global Perspective Bias | Biases in representing diverse global perspectives |
| Bisbee et al. [3] | Demographic Prediction Bias | Potential biases in demographic predictions using synthetic data |
| Kim and Lee [9] | Opinion Underrepresentation | Underrepresentation of certain opinions |
| Chu et al. [4] | Content Representation Bias | Biases in media content representation affecting predictions |
| Hwang et al. [7] | Demographic Bias | Reinforcement of existing societal biases |
| Sanders et al. [14] | Ideological Bias | Variability in effectiveness across demographic groups |
| Lee et al. [10] | Conditioning Bias | Importance of detailed conditioning to reduce biases |
| Sun et al. [17] | Response Bias | Harmlessness bias in responses to sensitive topics |

- Sun et al. [17] employed KL Divergence to quantify response biases particularly focusing on the tendency for responses to topics to downplay harm.
- Hwang et al. [7] utilized Pearson Correlation to examine biases illustrating how LLMs can perpetuate existing biases.
- Durmus et al. [5] applied Jensen Shannon Distance to gauge biases in representing perspectives.
- Sanders et al. [14] used MAPE to analyze biases revealing variations in effectiveness across demographic groups.
- Santurkar et al. [15] introduced Consistency Score and Representativeness Score as metrics for assessing discrepancies, between LLM generated opinions and actual viewpoints.

## 4.3 Performance and Potential

Studies, by Hwang et al. [7], Kim and Lee [9] have shown that LLMs can accurately predict opinion indicating a level of reliability. The use of datasets like GlobalOpinionQA [5] demonstrates the ability of LLMs to handle perspectives although their performance may vary among different demographic groups.

## 4.4 Challenges and Limitations

Various studies have identified biases in model predictions, including ideological biases, highlighting the need for effective bias mitigation strategies [2, 5]. The sensitivity to changes in data and prompts also emphasizes the importance of establishing methodologies for training and evaluating LLMs [3].

## 4.5 Ethical Considerations

Ethical concerns surrounding the reinforcement of existing biases are crucial when considering the use of LLMs. The potential for

these models to create echo chambers or amplify perspectives underscores the importance of implementing fairness and transparency measures in their deployment [10].

## 5 Conclusion

This analysis of LLMs in opinion survey studies reveals promising outcomes in terms of accuracy rates and strong correlations. However the existence of biases and ethical concerns underscores the need for research endeavors to improve bias mitigation strategies standardize assessment procedures and address implications to optimize the impact of LLMs on shaping opinion.

Our study suggests that while LLMs hold promise in revolutionizing opinion surveys it is essential to account for their limitations and ethical ramifications. By addressing these challenges LLMs have the potential to become tools, for academics, professionals and decision makers in understanding and predicting sentiment trends.

## References


[1] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. arXiv:2208.10264 [cs.CL] https://arxiv.org/abs/2208.10264

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. https://doi.org/10.1017/pan.2023.2 arXiv:2209.06899 [cs].

[3] James Bisbee, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. 2023. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. https://doi.org/10.31235/osf.io/5ecfa

[4] Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language Models Trained on Media Diets Can Predict Public Opinion. http://arxiv.org/abs/2303.16779 arXiv:2303.16779 [cs].

[5] Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. http://arxiv.org/abs/2306.16388 arXiv:2306.16388 [cs].

[6] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. http://arxiv.org/abs/2305.08283 arXiv:2305.08283 [cs].

[7] EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning Language Models to User Opinions. http://arxiv.org/abs/2305.14929 arXiv:2305.14929 [cs].

[8] Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal* 4 (2023), 100020. https://doi.org/10.1016/j.nlp.2023.100020

[9] Junsol Kim and Byungkyu Lee. 2024. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. http://arxiv.org/abs/2305.09620 arXiv:2305.09620 [cs].

[10] S. Lee, T. Q. Peng, M. H. Goldberg, S. A. Rosenthal, J. E. Kotcher, E. W. Maibach, and A. Leiserowitz. 2024. Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias. http://arxiv.org/abs/2311.00217 arXiv:2311.00217 [cs].

[11] Kiwan Maeng, Alexei Colin, and Brandon Lucia. 2017. Alpaca: intermittent execution without checkpoints. *Proc. ACM Program. Lang.* 1, OOPSLA, Article 96 (oct 2017), 30 pages. https://doi.org/10.1145/3133920

[12] Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics* 11, 1 (2024), 20531680241231468. https://doi.org/10.1177/20531680241231468 arXiv:https://doi.org/10.1177/20531680241231468

[13] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. arXiv:2402.06196 [cs.CL] https://arxiv.org/abs/2402.06196

[14] Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the Potential of AI-based Political Issue Polling. http://arxiv.org/abs/2307.04781 arXiv:2307.04781 [cs].

[15] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? http://arxiv.org/abs/2303.17548 arXiv:2303.17548 [cs].

[16] Gabriel Simmons and Christopher Hare. 2023. Large Language Models as Subpopulation Representative Models: A Review. http://arxiv.org/abs/2310.17888 arXiv:2310.17888 [cs].

[17] Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. 2024. Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information. http://arxiv.org/abs/2402.18144 arXiv:2402.18144 [cs].

[18] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient Large Language Models: A Survey. *OpenReview* (May 2024). https://openreview.net/forum?id=bsCCJHbO8A

[19] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (03 2024), 237–291. https://doi.org/10.1162/coli_a_00502 arXiv:https://direct.mit.edu/coli/article-pdf/50/1/237/2367175/coli_a_00502.pdf


Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009