Simulating Social Fingerprints: A DIF-Based Analysis of Structural Bias in LLM Responses

Anonymous ACL submission

Abstract

Recent studies have explored using large language models (LLMs) as virtual respondents in survey research, with a key challenge being to evaluate how well they align with human responses. This study applies Item Response Theory (IRT) and Differential Item Functioning (DIF)—methods commonly used in human surveys—to analyze item-level bias in LLM-generated responses. IRT estimates a respondent's latent trait from their answer patterns, while DIF statistically examines whether groups with the same trait respond differently depending on demographic attributes. We constructed personas with various demographic characteristics and simulated their responses to items from the American National Election Studies (ANES). The results show that LLMs replicate human-like bias directions and rankings of influential attributes, but the strength of the bias is substantially amplified. We also observed signs of social desirability bias in LLM responses to race-related items. This study demonstrates that, in the context of persona-assigned LLMs participating in surveys, IRT and DIF analyses enable quantitative, attribute-level bias evaluation—offering a meaningful contribution to the study of human–LLM alignment.

1 Introduction

LLMs are increasingly being adopted for tasks that simulate or supplement human judgment, including survey response modeling, public opinion estimation, and policy evaluation (Argyle et al., 2023; Horton, 2023; Ziems et al., 2024). As their applications expand into social scientific domains, it becomes critical to ask whether these models merely mimic surface-level human patterns or also reflect deeper structures of societal bias.

In particular, *structural bias* refers to systematic differences in responses that emerge not from individual preferences alone but from underlying demographic factors such as ideology, religion, gender, and race (Johnson et al., 2011). In the social sciences, such biases have long been analyzed through IRT (Rasch, 1993; Lord, 1952) and DIF (Scheuneman, 1979; Holland and Wainer, 2012) analysis. These methods have been applied for decades in large-scale surveys such as the ANES, the World Values Survey (WVS), and the General Social Survey (GSS) (Angoff, 2012).

This study begins with the following questions: **Do LLMs reproduce structural biases found in human survey responses? And can such biases be systematically detected using psychometric tools like IRT and DIF analysis?**

To investigate this, we construct a diverse set of personas based on the ANES dataset, varying in demographic and ideological attributes. We simulate responses to real survey items using several latest LLMs. Each persona's latent political ideology is estimated using an IRT-based approach, followed by DIF analysis to examine whether individuals with similar ideological traits respond differently depending on their demographic group.

Our results show that most LLMs exhibit bias directions similar to those of humans with respect to attributes such as political ideology and religion, while attributes like marital status show no significant effect in either case. However, the degree of bias tends to be exaggerated across most attributes.

This study offers a novel application of DIF analysis to LLM simulation contexts, showing its potential for assessing structural bias and contributing to future research on human–LLM alignment.

2 Related Work

Recently, researchers have begun to apply these human-oriented bias analysis methods to **LLMs**. For example, Bai et al. (2024) adapted the Implicit Association Test (IAT) to uncover hidden biases in LLMs, revealing stereotypical patterns across race, gender, and religion. Hu et al. (2025) evaluated

whether LLMs exhibit identity-based in-group favoritism, and Potter et al. (2024) showed that many LLMs tend to favor liberal political candidates, particularly in the U.S. context.

Some studies, however, suggest that LLMs demonstrate biases differently from humans. For instance, they may show lower sensitivity to variations in question wording (Tjuatja et al., 2024), or fail to reflect deeper perceptual differences, even when persona prompts are used (Giorgi et al., 2024). Wang et al. (2025) caution that LLMs may oversimplify identity expression and reinforce stereotypes.

However, most existing studies do not apply psychometric methods such as IRT or DIF. Prior work has mainly focused on analyzing response distributions or surface-level keywords. Few have examined whether LLMs and humans, under matched ideological traits, exhibit structural differences at the item level.

Our study uses IRT to align human and LLM responses, then applies DIF analysis to examine demographic effects on specific items. This approach aims to investigate the extent to which LLMs reflect *structural social biases* similar to those found in human behavior.

3 Methods

3.1 Dataset and Experimental Setting

We utilizes the ANES(The American National Election Studies, 2021) dataset, which provides a wide range of demographic information while excluding personally identifiable location data. This makes it well-suited for analyzing item-level structural bias in human responses and for constructing LLM personas based on demographic attributes.

The dataset consists of responses from 8,280 real individuals with diverse demographic backgrounds, which were used to construct an equal number of simulation personas. Each persona includes demographic attributes such as gender, race, religion, and political ideology, and served as the basis for the subsequent survey response simulations.

To ensure compatibility with IRT analysis, we selected survey items that were likely to reflect political ideology and used a consistent Likert scale (1: oppose, 2: neutral, 3: support). Ultimately, five items were chosen for analysis, covering the topics of gun control, immigration policy, welfare policy, transgender policy, and racial policy.

The simulation was implemented in Python and conducted using a range of LLMs suitable for large-scale experiments, including gpt-3.5-turbo, gpt-4o-mini, Claude-3-Haiku, Meta-LLaMa-3.1-8B, Google Gemini-2.0, and Mistral-7B. All models were run with the temperature parameter fixed at 0.7. The cost per full simulation ranged from approximately $2 to $5, with a runtime of 15 to 20 hours depending on the model. To mitigate stochastic variability and ensure stable trends, each model was simulated twice, and the results from both runs were used for analysis.

A complete prompt incorporating the persona, survey question, and instruction is presented in Appendix A.

3.2 Estimating Latent Political Traits Using IRT

For each of the five selected items, latent trait scores (theta) were estimated separately for human and LLM data using IRT (Rasch, 1993; Lord, 1952) analysis. To ensure consistency in item polarity, responses were recoded in advance such that '1 = progressive', '2 = neutral', and '3 = conservative'. The Graded Response Model (GRM) (Johnson et al., 2011; Van Der Linden and Hambleton, 1997) was employed to handle these polytomous responses.

$$P_{ijk} = \frac{1}{1 + \exp[-a_j(\theta_i - b_{jk})]} \quad (1)$$

Equation (1) defines P_{ijk} as the probability that respondent i selects category k or higher on item j. In this formulation, θ_i denotes the respondent's latent ideological trait, a_j is the item discrimination parameter, and b_{jk} is the threshold for category k on item j.

The estimated θ values were subsequently used as a reference point for comparing bias across demographic attributes in the DIF analysis.

3.3 DIF-Based Item-Level Bias Analysis

Based on the previously estimated latent political trait scores (θ), we conducted a DIF (Scheuneman, 1979; Holland and Wainer, 2012) analysis to assess whether individuals with similar ideological orientations respond systematically differently to survey items depending on their demographic attributes. For this analysis, we employed binary logistic regression to detect differential item functioning.

$$\text{logit}(P(Y=1)) = \beta_0 + \beta_1\theta + \beta_2\text{group}_1 + \beta_3\text{group}_2 + \cdots \quad (2)$$

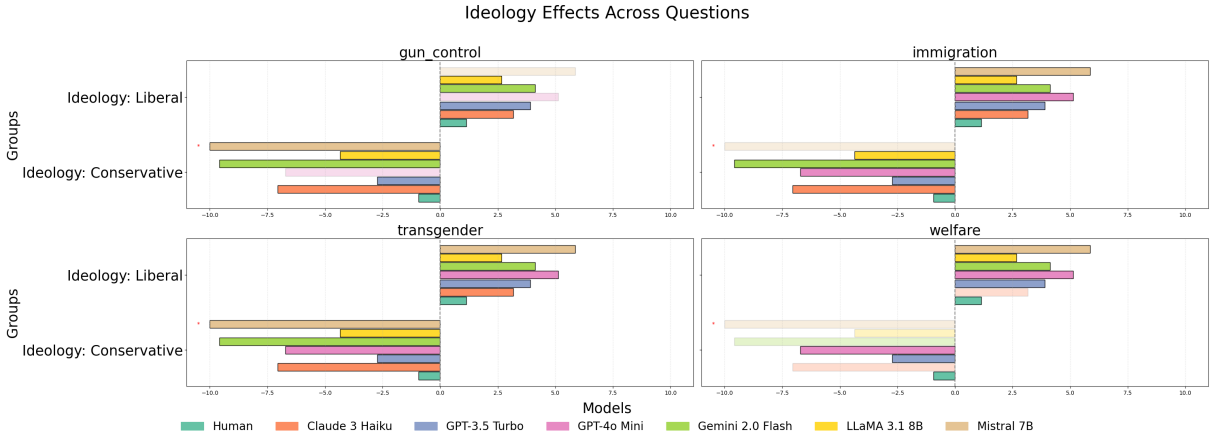


Figure 1: Each plot presents item-level bias coefficients associated with the political ideology attribute. Non-significant coefficients are shown with transparency. Values exceeding the clipping threshold of ±10 are capped at 10 and marked with an asterisk.

Rank	Human	GPT-3.5 Turbo	GPT-4o Mini	Claude 3 Haiku	Gemini 2.0 Flash	LLaMA 3.1 8B	Mistral 7B
1	**Ideology** (1.12)	**Ideology** (3.00)	**Ideology** (5.44)	**Ideology** (5.16)	**Ideology** (4.06)	**Ideology** (2.41)	**Ideology** (4.42)
2	**Religion** (0.60)	**Religion** (2.78)	**Religion** (2.84)	**Religion** (2.42)	**Religion** (2.85)	**Religion** (1.24)	**Religion** (2.17)
3	Education (0.48)	Gender (1.27)	Income (2.26)	Gender (2.00)	Income (1.46)	Gender (0.53)	Income (2.01)
4	Race (0.48)	Race (1.00)	Gender (2.19)	Education (1.49)	Education (1.35)	Race (0.49)	Gender (1.67)
5	Age (0.32)	Education (0.97)	Education (1.92)	Race (1.13)	Race (1.12)	Education (0.44)	Race (1.28)
6	Income (0.30)	Income (0.94)	Race (1.44)	Income (0.98)	Gender (1.12)	Income (0.44)	Education (0.88)
7	Gender (0.30)	Age (0.55)	Age (1.42)	Age (0.78)	Age (0.88)	Age (0.32)	Age (0.43)
8	**Marital*** (0.00)	**Marital*** (0.00)	**Marital*** (0.00)	**Marital*** (0.00)	**Marital*** (0.00)	**Marital*** (0.00)	**Marital*** (0.00)

Table 1: Demographic attributes ranked by average absolute DIF coefficients. Values in parentheses indicate effect size. Only attribute–item pairs significant in at least 6 of 7 models were included (35 total). **Marital*** was non-significant across all models but shown for reference.

Each item response was binarized (e.g., progressive choice = 1; neutral or conservative = 0). As shown in Equation (2), θ represents the latent ideological trait, and group_k refers to one-hot encoded demographic attributes. If the regression coefficient β_k associated with a specific demographic variable is statistically significant, the item is considered to exhibit differential response bias with respect to that attribute. For example, if respondents with the same latent ideological trait (θ) exhibit different response probabilities based on gender, this is interpreted as evidence of gender-related bias. Using this approach, we analyze which demographic attributes lead to structural bias at the item level in both human and LLM responses, as well as the direction of that bias (i.e., more progressive or more conservative).

4 Results & Discussion

4.1 Structural Bias Analysis

Figure 1 visualizes the regression coefficients by political ideology group (conservative/liberal) across items. The plots are based on the averaged results from two simulation runs. In each plot, the x-axis represents the direction of response probability, with negative values indicating more conservative responses and positive values indicating more progressive responses.

All LLM models exhibited bias directions generally aligned with human responses, but the magnitude of bias was substantially exaggerated. Notably, the Mistral 7B model reached the clipping threshold of ±10. While human respondents showed relatively moderate bias coefficients (e.g., –0.94 for conservative and +1.13 for progressive), LLMs produced extreme coefficients exceeding ±5.

These results indicate that LLMs are highly sensitive to the political ideology attributes embedded in the persona prompts and tend to exaggerate ideological differences. In addition to political ideology shown in Figure 1, LLMs also exhibited generally more extreme coefficients than humans for other attributes such as gender and religion. The corresponding visualizations are included in Appendix C.

Table 1 presents the average bias magnitude by attribute across models. In addition to political ideology, religion also showed 2–5 times larger coefficients in LLMs than in human data. Other

attributes, such as gender, income, education, race, and age, likewise exhibited 2–3 times greater coefficients in LLMs. These results suggest that while LLMs generally align with humans in directional trends, they tend to produce structurally exaggerated responses in terms of bias strength.

4.2 Similarity to Human Bias Patterns

The top- and bottom-ranked attributes shown in Table 1 exhibit similar patterns across both humans and LLMs. Ideology was the most influential attribute at the item level, consistently showing the strongest effect in both human and LLM responses, followed by religion. In contrast, marital status was consistently the least influential attribute for both humans and LLMs. This aligns with the common understanding that political ideology and religious beliefs are the most influential factors in human responses to the selected items (e.g., welfare policy, transgender-related policy), and demonstrates that LLMs tend to mimic this pattern as well.

Table 2 presents a comparison between the direction of item–attribute bias observed in humans (with negative values indicating conservative tendencies and positive values indicating progressive tendencies) and that observed in LLMs. Most LLM models exhibited a high level of agreement with human bias directions, ranging from **72%** to **82%**, with an average alignment rate of **77%**. This suggests that LLMs responded in a similar direction to humans on the majority of items. These results indicate that LLMs are capable of partially reproducing human-like bias structures at the item level.

Model	n_{same}	n_{total}	Alignment Rate
Gemini 2.0 Flash	106	130	0.815
GPT-3.5 Turbo	92	114	0.807
GPT-4o Mini	94	118	0.797
Claude 3 Haiku	82	108	0.759
LLaMA 3.1 8B	80	106	0.755
Mistral 7B	74	102	0.725
LLM Average	**528**	**678**	**0.776**

Table 2: Directional alignment rates between human and LLM responses. Alignment is defined as matching the sign of the human coefficient (positive = liberal, negative = conservative) for each item–group pair.

4.3 Social desirability bias in LLMs

For race-related items, most LLMs exhibited instability or highly skewed responses, often defaulting to extremely liberal or neutral positions. Due to this imbalance, regression analysis became infeasible for these items, and they were ultimately excluded from the DIF results. This may be attributed to alignment constraints designed to suppress potentially sensitive or controversial outputs in race-related contexts.

Such anomalies diverge significantly from human response patterns and highlight a key limitation of using LLMs as experimental agents in social science contexts.

4.4 Implications for Human–LLM Alignment in Survey Contexts

Ensuring the alignment between LLMs and humans is a critical challenge in terms of response reliability when LLMs are used as participants in survey research. This study demonstrates that, in the context where persona-assigned LLMs participate in surveys, it is possible to conduct quantitative, attribute-level bias analysis using IRT and DIF methods—an approach that, to the best of our knowledge, is the first of its kind.

Furthermore, this analysis enables comparison of which human attributes LLMs are aligned with or not, and is expected to make a meaningful contribution to the study of human–LLM alignment.

5 Conclusion

This study applied IRT and DIF analyses to persona-based LLM simulations, demonstrating that structural bias between humans and LLMs can be quantitatively assessed at the item level. The results showed that while LLMs exhibited partially similar directional patterns of bias to humans, the magnitude of these biases was often excessively amplified. In particular, for sensitive items, LLM responses differed significantly from those of human respondents, indicating potential limitations in using LLMs as participants in social science experiments. Moreover, this study illustrates the feasibility of using this approach to compare, at the item level, which demographic attributes LLMs are aligned with—or not—offering a meaningful contribution to research on human–LLM alignment.

6 Limitation

This study is based on simulation results derived from specific survey items and prompt conditions, and the interpretation should be considered within this experimental context. The number of items and the range of LLM models used were limited, making it difficult to generalize to broader policy issues or model architectures.

References

William H. Angoff. 2012. Perspectives on differential item functioning methodology. In *Differential item functioning*, pages 3–23. Routledge.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Xuechen Bai, Alex Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *Preprint*, arXiv:arXiv:2402.04105.

Salvatore Giorgi, Tianlin Liu, Abhisek Aich, Kirill Isman, Garrick Sherman, Zachary Fried, and Barrett Curtis. 2024. Modeling human subjectivity in llms using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188.

Paul W Holland and Howard Wainer. 2012. *Differential item functioning*. Routledge.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Tongtong Hu, Yevheniia Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Timothy P Johnson, Sharon Shavitt, and Allyson L Holbrook. 2011. Survey response styles across cultures.

Frederic Lord. 1952. A theory of test scores. *Psychometric monographs*.

Yunhan Potter, Sarah Lai, Juyeon Kim, Jacob Evans, and Dongwon Song. 2024. Hidden persuaders: Llms' political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275.

Georg Rasch. 1993. *Probabilistic models for some intelligence and attainment tests.* ERIC.

Janice Scheuneman. 1979. A method of assessing bias in test items. *Journal of Educational Measurement*, pages 143–152.

The American National Election Studies. 2021. The anes 2020 time series study. https://electionstudies.org/data-center/2020-time-series-study/. Accessed May 19, 2025.

Layne Tjuatja, Victoria Chen, Tony Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Wim J Van Der Linden and Ronald K Hambleton. 1997. Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory*, pages 1–28. Springer.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

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):237–291.

Appendix: Code Availability

An anonymous implementation of the simulation code and data processing scripts is available at the following URL:

https://anonymous.4open.science/r/project-anon-C033/

This repository has been anonymized in accordance with the ACL reviewing policy and will remain accessible during the review process.

A Prompts Used in the Simulation

We present an example prompt combining persona, question, and instruction. This format was used in all simulation trials.

```
I am 46 years old, male, asian, in the
income bracket '$175,000-249,999',
with an education level of bachelor's
degree, who identifies as conservative,
and religiously identifies as something
else.

When asked the following question, I
respond based on my beliefs and background.

Should the federal government make it more
difficult for people to buy a gun?
1. More difficult
2. Easier
3. Keep the rules about the same
Please respond with only the number (1 to
3).
```

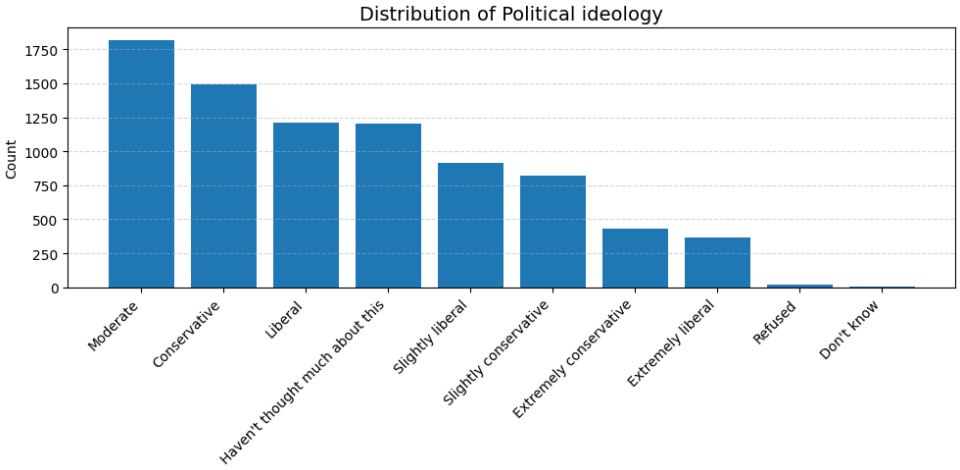
B Demographic Distribution of ANES Personas

We provide summary statistics of the demographic attributes used to construct ANES-based personas.

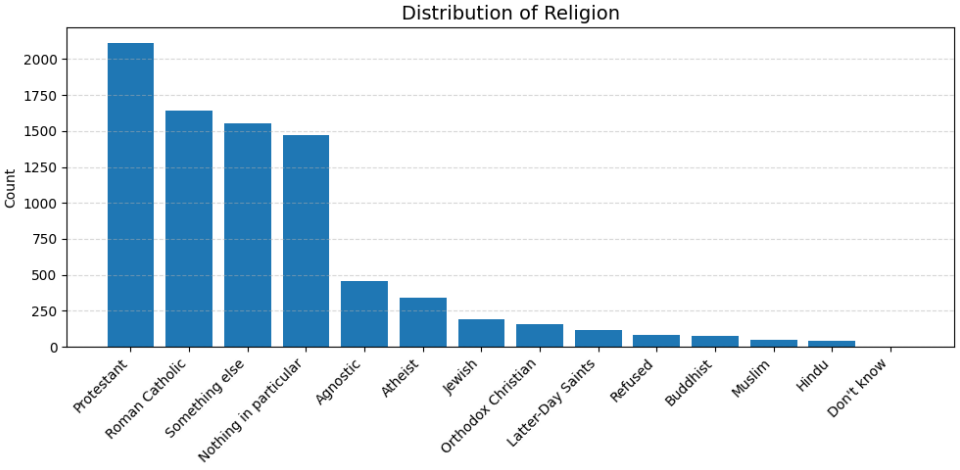
These include political ideology, religion, race, income, gender, marital status, and education. The distribution plots in Figure 2 show the diversity of the sampled population across these variables.

C Visualization: DIF Analysis

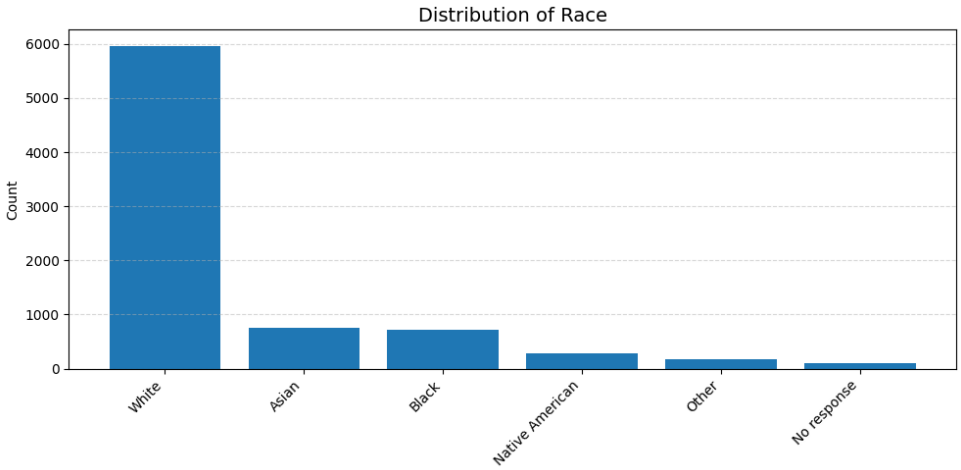
We visualize the estimated effect sizes from the DIF analysis across demographic groups. Figure 1 displays the regression coefficients by group, showing the direction and significance of demographic influence on each survey item.


Distribution of Political ideology
Count
0
250
500
750
1000
1250
1500
1750
Moderate
Conservative
Liberal
Haven't thought much about this
Slightly liberal
Slightly conservative
Extremely conservative
Extremely liberal
Refused
Don't know

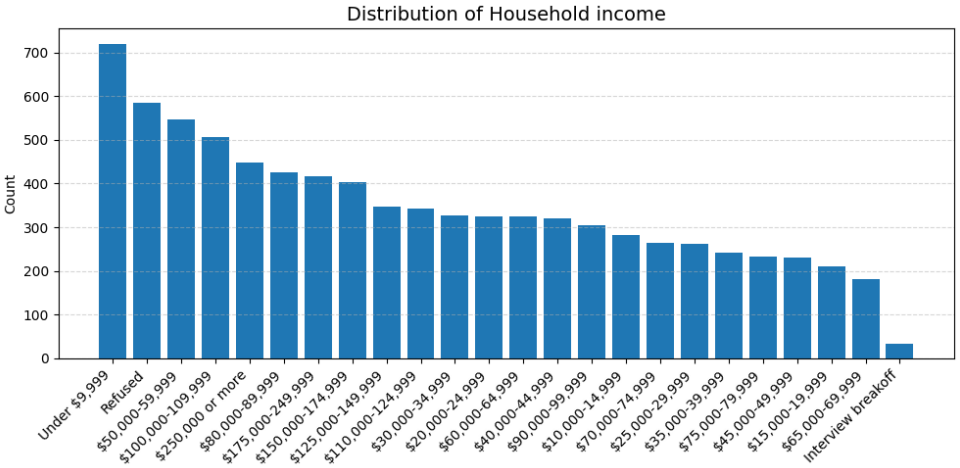
(a)


Distribution of Religion
Count
0
250
500
750
1000
1250
1500
1750
2000
Protestant
Roman Catholic
Something else
Nothing in particular
Agnostic
Atheist
Jewish
Orthodox Christian
Latter-Day Saints
Refused
Buddhist
Muslim
Hindu
Don't know

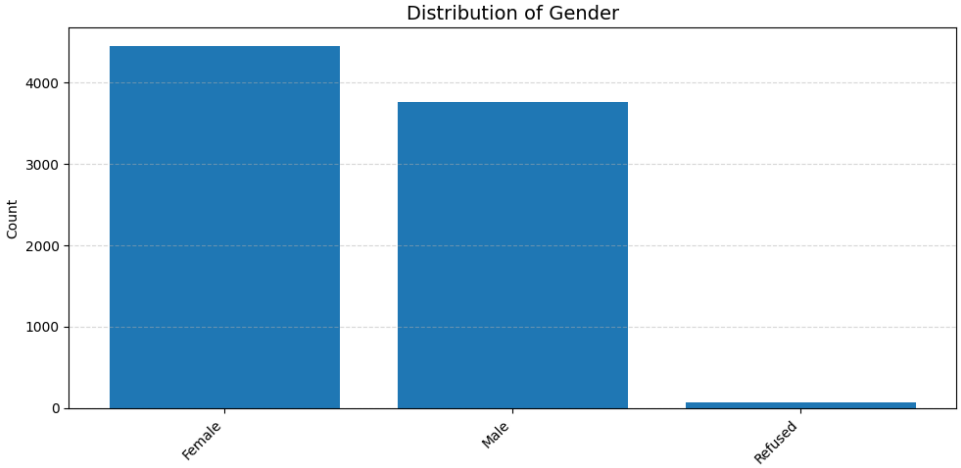
(b)


Distribution of Race
Count
0
1000
2000
3000
4000
5000
6000
White
Asian
Black
Native American
Other
No response

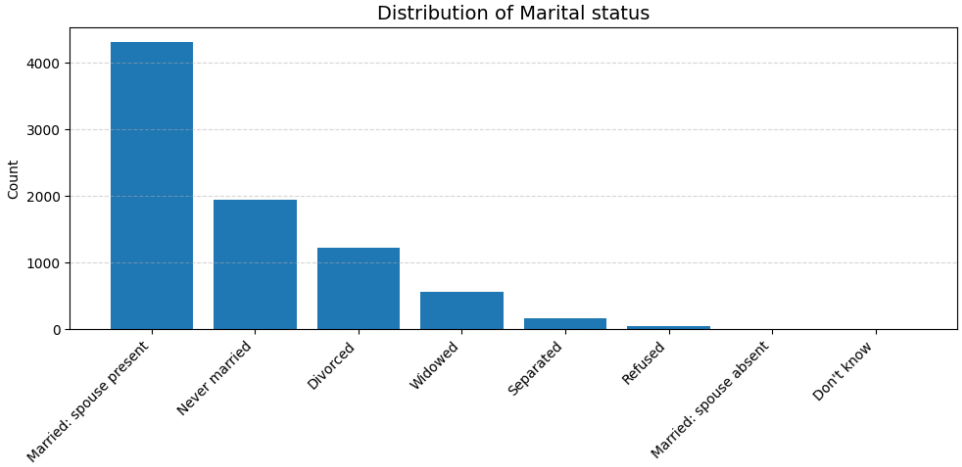
(c)


Distribution of Household income
Count
0
100
200
300
400
500
600
700
Under $9,999
Refused
$50,000-59,999
$100,000-109,999
$250,000 or more
$80,000-89,999
$175,000-249,999
$150,000-174,999
$125,000-149,999
$110,000-124,999
$30,000-34,999
$20,000-24,999
$60,000-64,999
$40,000-44,999
$90,000-99,999
$10,000-14,999
$70,000-74,999
$25,000-29,999
$35,000-39,999
$75,000-79,999
$45,000-49,999
$15,000-19,999
$65,000-69,999
Interview breakoff

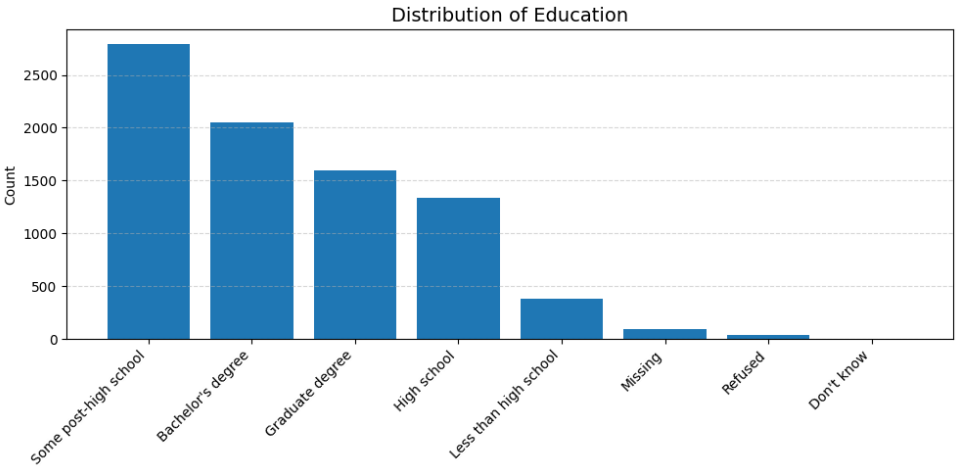
(d)


Distribution of Gender
Count
0
1000
2000
3000
4000
Female
Male
Refused

(e)

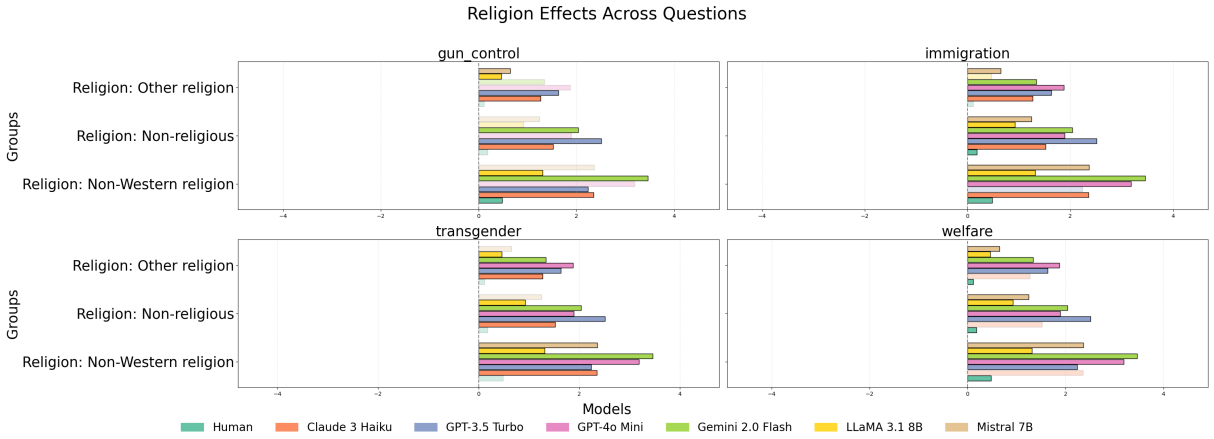

Distribution of Marital status
Count
0
1000
2000
3000
4000
Married: spouse present
Never married
Divorced
Widowed
Separated
Refused
Married: spouse absent
Don't know

(f)

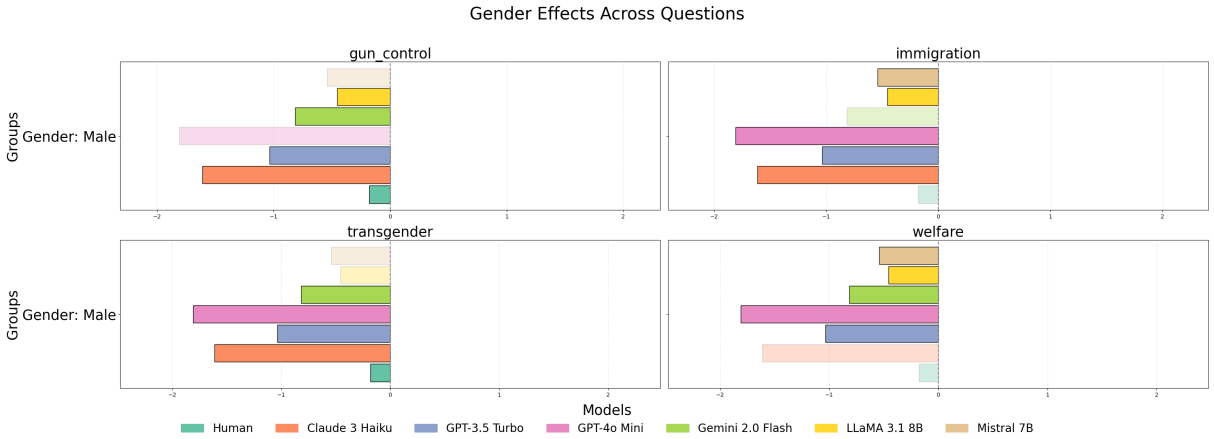


(g)

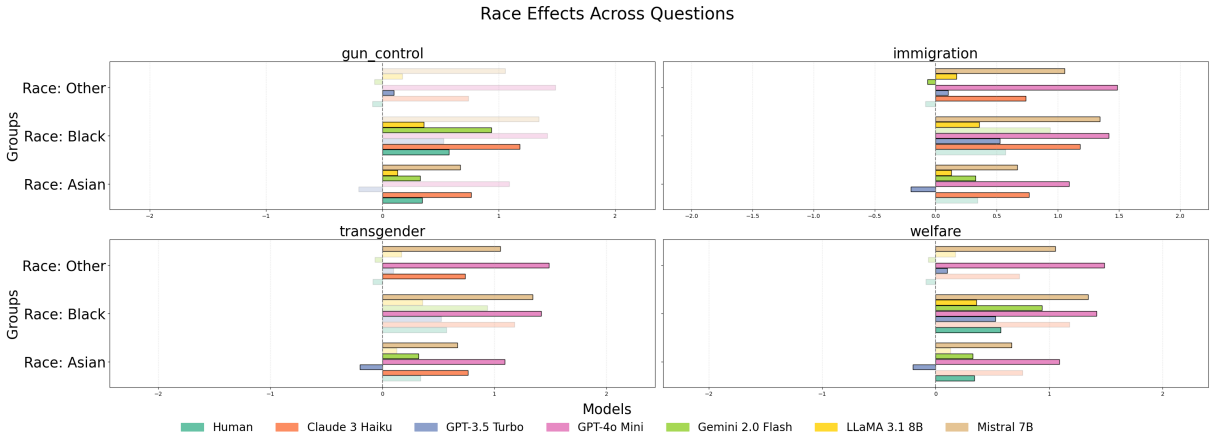
Figure 2: Distribution of demographic characteristics among ANES-based personas used in the simulation study. Each subfigure presents the frequency counts for a specific variable: (a) Political ideology, (b) Religion, (c) Race, (d) Income, (e) Gender, (f) Marital status, and (g) Education. These distributions reflect the diversity of the underlying ANES data and were used to construct the persona pool.


Religion Effects Across Questions
gun_control
immigration
transgender
welfare
Groups
Religion: Other religion
Religion: Non-religious
Religion: Non-Western religion
Models
Human
Claude 3 Haiku
GPT-3.5 Turbo
GPT-4o Mini
Gemini 2.0 Flash
LLaMA 3.1 8B
Mistral 7B

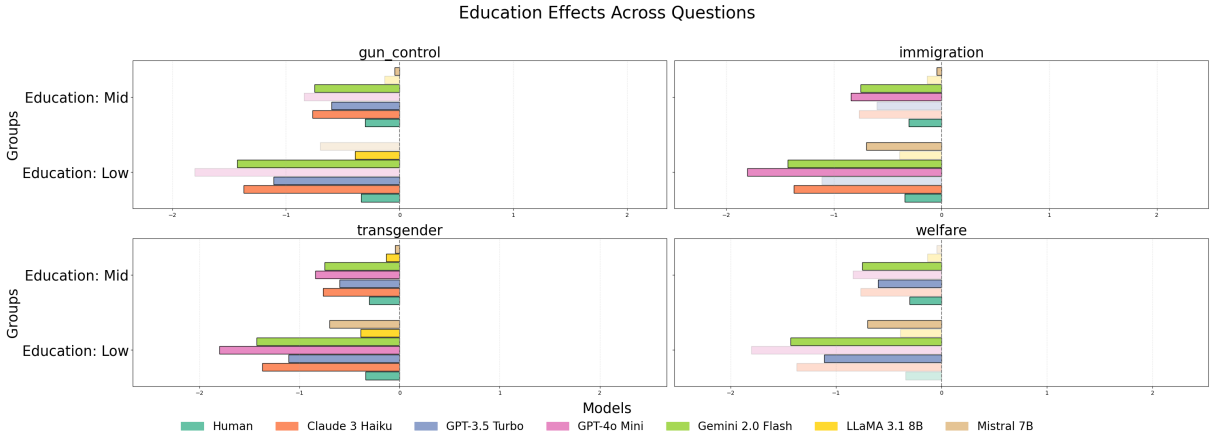
(a)


Gender Effects Across Questions
gun_control
immigration
transgender
welfare
Groups
Gender: Male
Models
Human
Claude 3 Haiku
GPT-3.5 Turbo
GPT-4o Mini
Gemini 2.0 Flash
LLaMA 3.1 8B
Mistral 7B

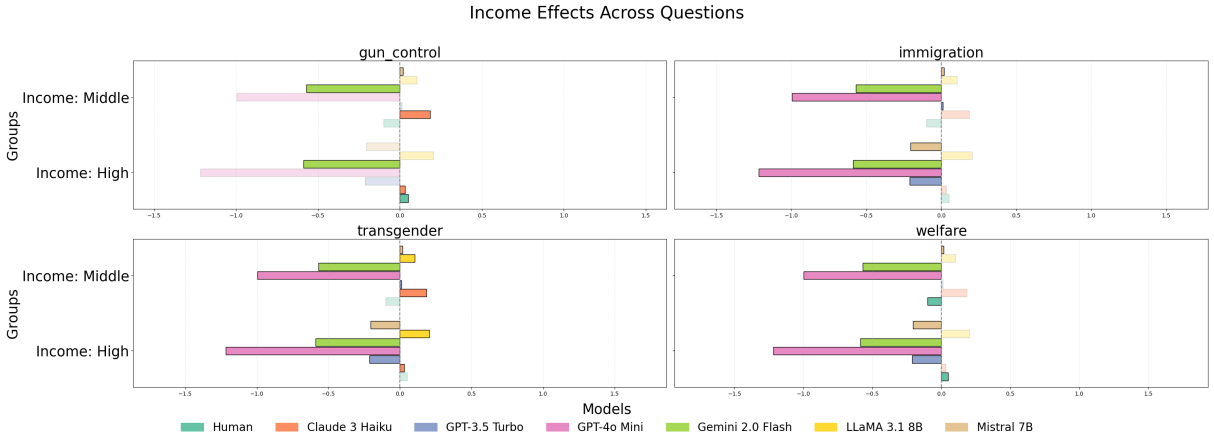
(b)


Race Effects Across Questions
gun_control
immigration
transgender
welfare
Groups
Race: Other
Race: Black
Race: Asian
Models
Human
Claude 3 Haiku
GPT-3.5 Turbo
GPT-4o Mini
Gemini 2.0 Flash
LLaMA 3.1 8B
Mistral 7B

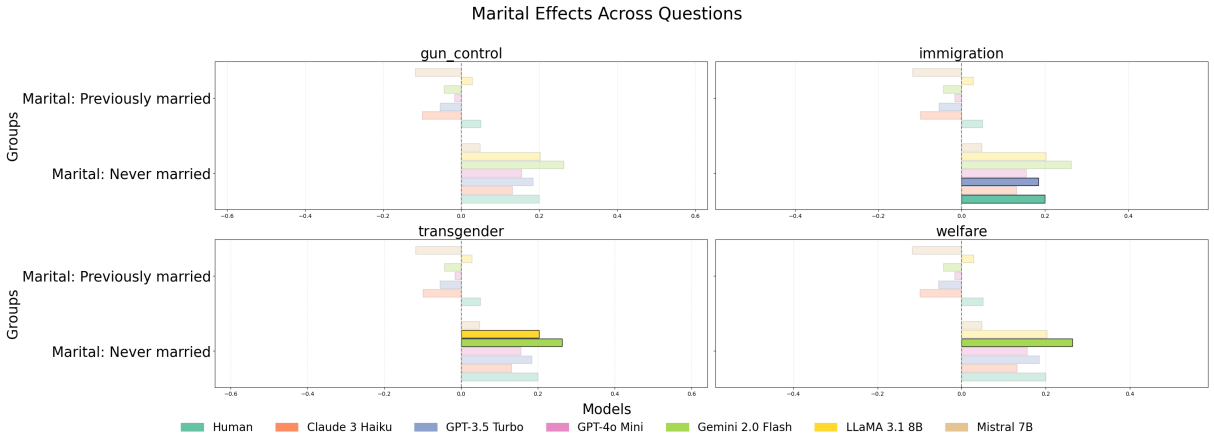
(c)



(d)



(e)



(f)

Figure 3: Directional DIF coefficients by demographic variable. Each subfigure (a–f) presents the regression coefficients for a specific group variable across survey items. Bars indicate the direction and magnitude of group effects (positive = progressive, negative = conservative), and transparency reflects statistical significance (opaque = significant at $p < .05$). (a) Religion, (b) Gender, (c) Race, (d) Education, (e) Income, and (f) Marital status.