The Political Nature of Conversational LLMs Using Item Response Theory

Kaisen Krishnamurthy Andrew B. Hall

RESEARCH QUESTION

Where do Conversational Large Language Models (LLMs) fall on the political spectrum? How are their political opinions influenced by user input?

1. MOTIVATION

As LLMs become increasingly integrated into general discourse...

- Politically biased LLMs could exacerbate polarization
- LLMs that reflect user preferences may create echo chambers, also potentially exacerbating polarization.

2. DATA

- We use a CES Stanford dataset containing 133 binary political questions with 1300 human respondents (2010)
- We prompted a set of LLMs with these 133 modified binary political questions
- Questions are modified by attaching a prepend and postpend to increase the probability of the LLM returning a response

3. METHODOLOGY

We use a 2-parameter item response theory model (IRT) to measure the political sentiment behind LLMs set of responses, with the CES dataset creating the ideological distribution, outlined by Fowler et al (2023) and Clinton, Jackman, and Rivers (2004)

$$P(y_{ij} = 1|x_i, \alpha_j, \beta_j) = \Lambda(\beta_j(x_i - \alpha_j))$$

- y_{ij} indexes the binary question response (1 is the positive response, 0 is the negative response)
- i indexes the respondents
- j indexes the questions
- x indexes the latent ideal point, translating to a place on the political spectrum
- α indexes the cutpoint parameter, question specific
- β indexes the discrimination parameter, question specific
- Λ is the CDF

The likelihood for a respondent's answers across all questions can be written as such

$$L(y_i; \alpha, \beta) \approx \sum_{k=1}^{M} \prod_{j \in J_i} \Lambda(\beta_j(x_k - \alpha_j))^{y_{ij}} (1 - \Lambda(\beta_j(x_k - \alpha_j)))^{1-y_{ij}}$$

- x_k are sampled ideal points sing Monte Carlo sampling from distribution $f_n(x)$ given the nth iteration.
- We find α_n and β_n by maximizing the log likelihood function:

$$\sum_i \log L(y_i; \alpha, \beta)$$

- Given an α_n and β_n, we find a new $f_{n+1}(x)$ and iterate until convergence.

- LLMs are either asked these political questions 10 times, or if available, token probabilities are gathered. Their responses are converted to a probability of the model returning the positive response.
- Once α and β values have been acquired, LLM x values, indexing political sentiment, can be solved for by using the initial equation.

6. CONCLUSIONS

1. This analysis provides substantial evidence that with default prompting, LLMs demonstrate significant left-leaning political bias. We can conclude that most mainstream AI systems reflect bias in one ideological direction.

- This could split the population into two groups: Those aligning with the viewpoints of AI systems will trust the AI-generated content as factual, while those who do not align with these viewpoints will view AI systems as tools of political propaganda.

2. The second step of this analysis provides substantial evidence that LLMs are highly reactive to user context, reflecting ('echoing') the user's opinions.

- This suggests that users that have interacted with LLMs at a significant volume are currently interacting in political echo-chambers. This could also further the political divide: Users will have their existing biases reinforced, resulting in heightened polarization.

7. LIMITATIONS & NEXT STEPS

- LLMs varied in the number of questions they answered. The issue with this lies in the potential significance of non-responses. If an AI system consistently avoids a particular type of question, it will influence the overall results.
- For convenience, we have been using a 2010 dataset, but we plan to test more recent datasets.
- It is not guaranteed that the biases observed in LLM's responses to political questions will manifest for users interacting with these models in a typical setting.

4. RESULTS

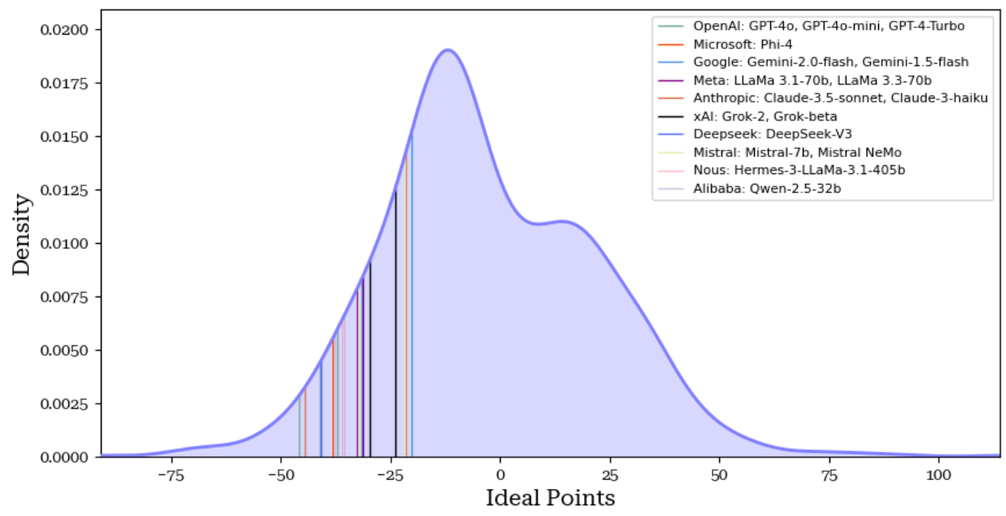

Figure 1. KDE Plot of Ideal Points (left-right on the political spectrum)

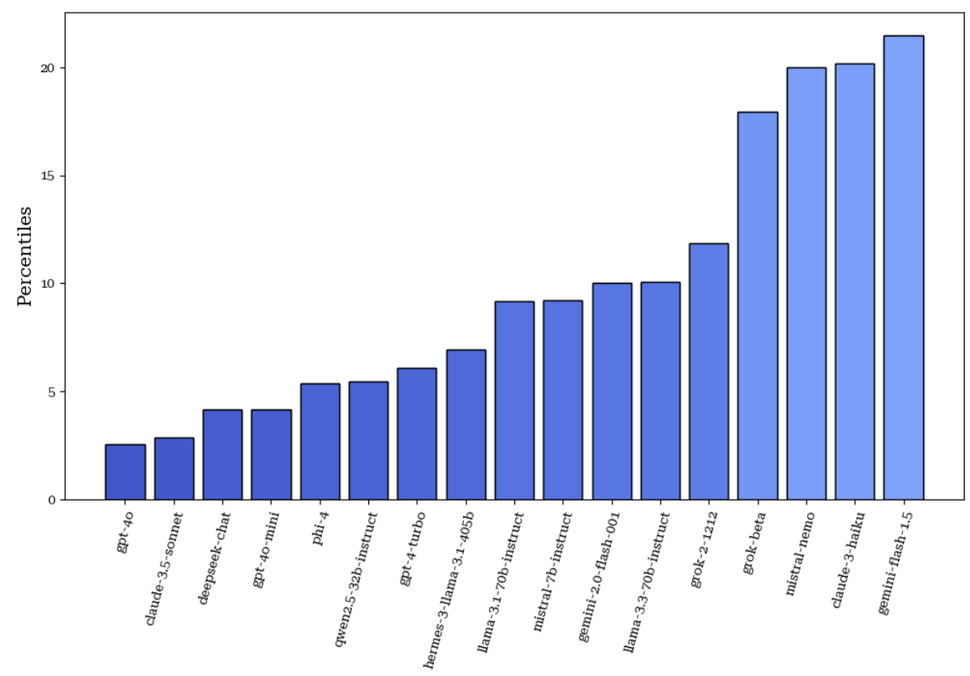

Figure 2. Percentile of LLMs Relative To Human Respondents

5. ECHO

- The concept of LLM "echo" refers to an LLM that, given sufficient user context, adopts positions that closely align with those emphasized by the user (Perez et al., 2022).
- At a given point on the political spectrum shown in Figure 1, we take a human respondents dataset and ask GPT-4.0-mini to summarize the political ideology of that respondent
- We take the summary, attach it to the modified question outlined previously, and input this new set of prompts into an LLM

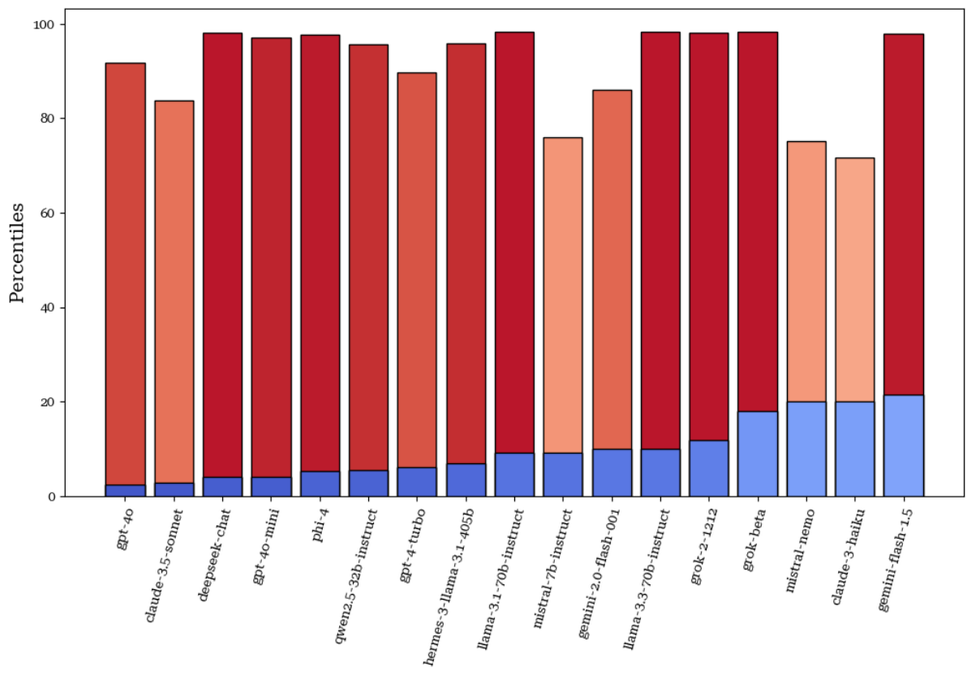

Figure 3. Percentile on the Political Spectrum With Context From a 100th Percentile Conservative

8. WHAT DO WE ADD?

Many existing studies measuring political bias have subjected LLMs to political surveys. These lack direct comparability to the human political distribution. We have developed a methodology that directly compares LLMs political ideology to that of human respondents, returning a percentile value on the human political spectrum rather than an arbitrary score. Additionally, we use this method to quantify echoic behavior in AI systems.