Harvard Data Science Review • Issue 5.4, Fall 2023

Demonstrations of the Potential of AI-based Political Issue Polling

Nathan E. Sanders[1] Alex Ulinich[2] Bruce Schneier[3]

[1]Berkman Klein Center, Harvard University, Cambridge, Massachusetts, United States of America,
[2]Mountain View High School, Mountain View, California, United States of America,
[3]Harvard Kennedy School, Harvard University, Cambridge, Massachusetts, United States of America

The MIT Press

Published on: Oct 27, 2023

DOI: https://doi.org/10.1162/99608f92.1d3cf75d

ABSTRACT

Political polling is a multi-billion-dollar industry with outsized influence on the societal trajectory of the United States and nations around the world. However, in recent years it has been severely challenged by rising nonresponse rates and other factors that stress its cost, availability, and accuracy. At the same time, artificial intelligence (AI) chatbots such as ChatGPT have become highly compelling stand-ins for a wide range of human behavior, powered by increasingly sophisticated large language models (LLMs). Because these LLMs are trained on huge corpora of writing by diverse people captured from across the Internet, they are potentially capable of representing a wide range of beliefs on many policy issues. Could AI chatbots be an effective tool for anticipating public opinion on controversial issues to the extent that they could be used by campaigns, interest groups, and polling firms?

We have developed a prompt engineering methodology for eliciting humanlike survey responses from ChatGPT, which simulate the response to a policy question of a person described by a set of demographic and ideological factors, and produce both an ordinal numeric response score and a textual justification. We execute large-scale experiments using this method, querying GPT for thousands of simulated responses at a cost that is at least three orders of magnitude lower than human surveys. We compare this simulated data to human issue polling data from the Cooperative Election Study (CES).

We find that ChatGPT is sometimes effective at anticipating both the mean level and distribution (correlation $> 85\%$, normalized Earth mover's distance $\lesssim 0.2$) of public opinion on a variety of policy issues well established in its training data such as abortion bans and approval of the US Supreme Court, particularly in their breakdown along ideological lines. However, it is much less successful at anticipating demographic (age, race, and gender) differences between respondents. Moreover, ChatGPT tends to overgeneralize its conception of ideological differences to new policy issues that arose after its training data was collected, such as American support for involvement in the war in Ukraine. Our work has implications for our understanding of the strengths and limitations of the current generation of AI chatbots as virtual publics or online listening platforms, future directions for LLM development, and applications of AI tools to the political domain.

Keywords: machine learning, artificial intelligence, public polling, large language models

Media Summary

Could AI tools like ChatGPT supplement political polling, focus groups, online social listening tools, or market research studies? This research tests the ability of ChatGPT to generate synthetic survey responses to issue polling questions and compares those responses to surveys of real humans. ChatGPT proves to be successful at responding like real Americans to some questions with a strong partisan divide, but it often fails to anticipate differences in public opinion along other human dimensions, such as demographics like age, race, and gender. ChatGPT also goes too far in extrapolating the expected partisan differences in response to events that took place after the training data it was created with, such as the breakout of the war in Ukraine. With further development to improve the range of policy views and demographic trends that they can accurately reproduce, such as ways to incorporate data on current events and systems to mitigate the potential for bias to be imprinted from their training data, AI chatbot systems could become a useful tool to political campaigns, interest groups, and pollsters.

---

1. Introductions

While survey experiments and polling have been powerful tools for political campaigns, parties, and advocacy organizations in the United States and around the world for centuries (Splichal, 2022), in recent years the cost and difficulty of operating polls has grown dramatically. Political polling firms commonly recruit panels intended to be representative of, and to achieve high coverage of, their targeted population, such as eligible voters nationally or likely voters in a voting district. Reaching these populations has become harder primarily because of the growth in survey nonresponse internationally: the failure to contact or refusal of potential participants to be surveyed due to factors such as lack of time, disinterest, and distrust (Luiten et al., 2020). Moreover, the migration of respondents to new technologies such as cell phones and the Internet, which have uneven and evolving penetration and usage across regions and demographic groups, has constrained the coverage of survey samples (Berinsky, 2017).

These effects have generated simultaneous challenges for the quality and cost of political polling, as biases in political engagement and hyperpolarization manifest on response rates (Cavari & Freedman, 2023; Olson et al., 2020). A vast literature has developed on statistical methodologies for designing and postprocessing survey data to overcome these challenges, including methods such as demographic weighting and poststratification (see e.g., Berinsky, 2017; Isakov & Kuriwaki, 2020; Kennedy et al., 2018). In particular, pollsters have explored methodologies that enable meaningful public opinion research from digital platforms such as Facebook and other social media platforms, where traditional techniques of probability sampling cannot be applied because of the lack of a conventional sampling frame and researcher-controlled contact mechanism (Murphy et al., 2014; Schneider & Harknett, 2022). These various methodologies seem to have been successful at maintaining the predictive accuracy of election polling thus far, even as nonresponse has proliferated (Jennings & Wlezien, 2018), and yet there is widespread interest in finding transformative new models for measuring public opinion that could lead to more cost-effective, sustainable, and more reliable polling results (Bailey, 2023; Concha, 2020; Graham, 2020; Kennedy et al., 2023; Montgomery, 2020; S. Roberts, 2020; Silver, 2021).

As statistical methodologies have come to play a critical role in collecting, processing, and interpreting political polling data, machine learning (ML) and artificial intelligence (AI) systems may further revolutionize this domain. In particular, large language models (LLMs) such as ChatGPT, which can be incorporated into AI chatbots and other systems capable of providing humanlike responses to natural language prompts, have a wide variety of potential applications in democratic processes, such as assisting lobbying firms (Nay, 2023;Sanders & Schneier, 2023a), helping citizens and stakeholders to formulate and advocate for their opinions (Schneier et al., 2023), facilitating connections between candidates and voters (Sanders & Schneier, 2023c), and even helping humans social engineer or hack political systems (Sanders & Schneier, 2021,2032b;Schneier, 2021). Already, researchers have experimented with a variety of social science research and public polling applications of LLMs, such as coding open-ended survey responses (Mellon et al., 2022), providing synthetic participants for human subjects research (Aher et al., 2023;Dillion et al., 2023), mimicking consumer responses to market factors like price sensitivity (Brand et al., 2023), inferring the ideology of a politician (Wu et al., 2023), representing the personality traits of psychological profiles (Jiang et al., 2023), simulating economic behavior (Horton, 2023), simulating feelings toward political parties and groups (Bisbee et al., 2023), and simulating election results (Argyle et al., 2023).

Because they are trained on wide Internet corpora including opinion writing from a diverse range of people, LLM's have a compelling ability to represent different perspectives and to perform a wide range of tasks without specialized training (Brown et al., 2020; Kojima et al., 2022; Agüera y Arcas, 2022). We therefore hypothesize that they may be effective at generating individualized responses to policy preference questions that can account for the same factors that influence human respondents, such as demographics and ideology.

However, the nature of LLMs limits their potential effectiveness as opinion sampling tools. Like social media platforms, AI chatbots do not have well-defined sample frames or well-understood coverage characteristics. Moreover, unlike true survey platforms, using LLMs does not actually involve any solicitation of opinion from an authentic human individual. Instead, LLMs generate a response predicted to be most acceptable to the user on the basis of a training process such as reinforcement learning with human feedback (Ziegler et al., 2019), which may therefore reflect the incomplete or biased properties of its training data set. Some specific biases of Internet corpora-trained LLMs are coming into focus. One study attempted to assess the age and gender characteristics of ChatGPT by prompting it to express a demographic profile, finding that its responses are biased toward a young (< 30 years old) and female profile (Miotto et al., 2022). Other investigators identified that an earlier model, GPT-2, is biased in its representation of the opinions of people from nations underrepresented in Internet usage (Narayanan Venkit et al., 2023). Regardless of their ability to reflect the perspectives of a given demographic group, AI models may also exhibit bias in the text they generate; for example, in an analysis of the Bidirectional Encoder Representations from Transformers (BERT) model, researchers found that neural embeddings learn harmful stereotypes about persons with disabilities (Hutchinson et al., 2020). For applications to survey research specifically, the ability to reproduce such biases and stereotypes may enable the LLM to authentically reflect the beliefs of the human population, even if inaccurate or morally negative; but only to the extent that its training data realistically reflects these sentiments.

In this work, we seek to test the capability of current generation AI tools to accurately reflect distributions of public opinion, and to expose insight into its effective sociodemographic coverage as a polling instrument, using a generally available LLM and real public opinion survey questionnaires. We have developed experimental methods (Methods, §2) to prompt the AI chatbot ChatGPT to generate public polling–like responses to evaluate how well it can simulate a survey panel. We test the model's ability to reflect the shift in valence between demographic and ideological groups across a variety of issues, as well as reasonably reproduce the key arguments appealed to by each subgroup (Results, §3). We provide an interpretation of this capability in the context of prior Internet-assisted approaches to public opinion research, discuss the limitations of this approach and the current generation of tools, and the implications these capabilities may have as they improve (Discussion, §4), before concluding (Conclusion, §5).

2. Methods

We explore the viability of AI language models to simulate public opinion polling responses by developing a system that automates querying an LLM based on the questionnaire of a survey previously given to people, so that the resulting AI responses are aligned and comparable to human data. The code and data associated with this paper has been published on GitHub.[1]

2.1. Large Language Model

We use the OpenAI Chat Completion API endpoint, through OpenAI's $openai$ Python library,[2] to query the $gpt\text{-}3.5\text{-}turbo\text{-}0301$ LLM for polling responses. This model was the most recent model from OpenAI optimized for chat applications and made generally available as of April 2023[3]; it is trained on data samples written as late as September 2021.[4] The GPT-3.5 model is a Generative Pretrained Transformer (GPT) language model extended with supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). OpenAI has not fully disclosed the training data used for the GPT-3.5 model, but, based on past publications, it is widely understood to include a wide range of text scraped from diverse web and book data sources, including the Common Crawl and Wikipedia (Brown et al., 2020).

We generate a balanced sample of $n = 100$ responses per prompt per demographic cross-tab per issue across ideology (in five bins) and three demographic fields with simple categorizations (age in four bins, 'man' or 'woman' gender, and 'White' or 'non-White' race), for a total of 8,000 responses across each of seven issue prompts (see Table 1) for 56,000 total responses. Note that this balanced sample does not, therefore, represent any particular target population such as US adults, as our focus is on understanding the performance of LLM's in representing the viewpoints within and across distinct subgroups. Because LLMs offer the opportunity to generate data for arbitrary sub-populations at arbitrary sizes, the process to generate a sample representative of a population with defined demographic characteristics is trivial, if the model is successful at accurately reproducing the views of each subgroup. This is similar to the common procedure of reweighting human survey response data sets (e.g. Ansolabehere & Rivers, 2013). Regarding our selected demographic classes, we acknowledge that binary categorizations for gender and race are reductive and far from representative of the full spectrum of human gender and racial identity. Our reason for focusing on these broad classes is to enable initial statistical comparisons with demographic groups well sampled in the CES data set. Future work should further explore the representation of AI-generated responses associated with nonbinary gender and more diverse racial identities.

Table 1. Survey question prompts used in ChatGPT simulation, adapted from the CES questionnaire.

Name	CES Code	Prompt text
$police\_safety$	`CC22_307`	How do the police make you feel?
$scotus\_approval$	`CC22_320c`	Do you approve of the way the US Supreme Court is doing its job?
$ukraine\_war$	`CC22_321_1`	The United States should not get involved in response to Russia's 2022 invasion of Ukraine
$prescription\_import$	`CC22_327d`	Thinking now about health care policy, would you support or oppose a policy to allow states to import prescription drugs from other countries
$abortion\_ban$	`CC22_332f`	On the topic of abortion, do you support or oppose a policy to make abortions illegal in all circumstances?
$increase\_fuel\_production$	`CC22_333e`	Increase fossil fuel production in the U.S. and boost exports of U.S. liquefied natural gas
$gun\_background\_checks$	`CC22_330e`	Improve background checks to give authorities time to check the juvenile and mental health records of any prospective gun buyer under the age of 21

These queries were executed at a cost of about $18 USD through the OpenAI API, whereas an online survey of 50,000+ responses on a human population would cost at least 1,000 times that much. Beyond the data collection cost, the AI polling method demonstrated here still requires human engagement to choose questions and implement prompts (similar to the effort required to develop a questionnaire for a human survey panel) and requires human engagement to review and interpret the results. While AI assistance may also reduce the effort required for questionnaire design and survey data analysis, we do not explore such methods in this work.

LLMs can be sensitive to the way questions are phrased and what information is provided to prime them before answering a question. We arrived at a prompt suitable for simulating public polling responses aligned to an established survey questionnaire through several iterations of trial and error in prompt engineering. We used the following prompt template when querying the LLM,

> Please write a 1-paragraph letter to the editor from the perspective of a $\{gender\}$ living in the United States in the age range of $\{age\}$ years who identifies as $\{white\}$ expressing a clear point of view on the policy proposal to: "$\{issue\}$." Before the letter, summarize their position with a 'Position score': statement followed by a single number (strictly numeric, with no other description) representing the person's position on the issue on a $\{cardinality\}$-point scale, where 1 represents the position "$\{low\_level\}$" and $\{cardinality\}$ represents the position "$\{high\_level\}$,"

where $\{gender\}$, $\{age\}$, and $\{white\}$ are demographic features; $\{issue\}$ represents the question text from a survey given to humans (§2.2); $\{cardinality\}$ is the maximum value of the numeric response scale; and $\{low\_level\}$ and $\{high\_level\}$ are descriptions of the bottom and top end of the response scale as defined in the polling questionnaire. The prompt component describing the 'Position score': successfully formats the output so that an ordinal numeric response value can be extracted from the plaintext completion with a simple regular expression. Additionally, we extract the textual descriptors of the top and bottom options on the original scale from the survey questionnaire to align the LLM outputs to the scale the human respondents used.

The prompt template defined above evolved significantly over the course of our experimentation. Initially, we did not include a "Position score" requirement in the prompt. We first tested the model's ability to generate realistic-seeming textual arguments in response to policy issue questions, from various demographically aligned points of view. Having initially vetted this capability, we then added a brief instruction to the prompt to assign a score on a 1–5 rating and verified that the generated ratings generally agreed with the textual letter generated by the model. However, we identified two further challenges: 1) the generated position score would be formatted inconsistently and was difficult to extract from the generated text without manual review and 2) the model would sometimes flip the polarity of the scoring scale, such that a given position would be variously represented as a score of 1 or 5. To address issue 1, we added far more explicit formatting instructions ("Before the letter, summarize their position with..."), which succeeded at enforcing a formatting for the score that could be reliably extracted. To address issue 2, we added explicit definitions to the template of the low and high position levels. In addition to ensuring a consistent directionality of the position score, this instruction has the added benefit of aligning the model-generated score to the rating scale used in the human survey. At each step, we iterated on the prompt design based on inspection of the GPT responses in isolation, and not on the basis of comparisons to human data, mitigating the possibility of overfitting the prompt design to the benchmark.

We use the default OpenAI system prompt of, “You are a helpful assistant.” The LLM is not prompted sequentially; rather, each query is executed as an individual user prompt immediately following the system prompt. Importantly, this allows us to simulate semi-independent poll responses, where the LLMs are probabilistically generating various isolated responses similar to sampling a human population.[5]

Figure 1 illustrates an example survey response distribution from this method.

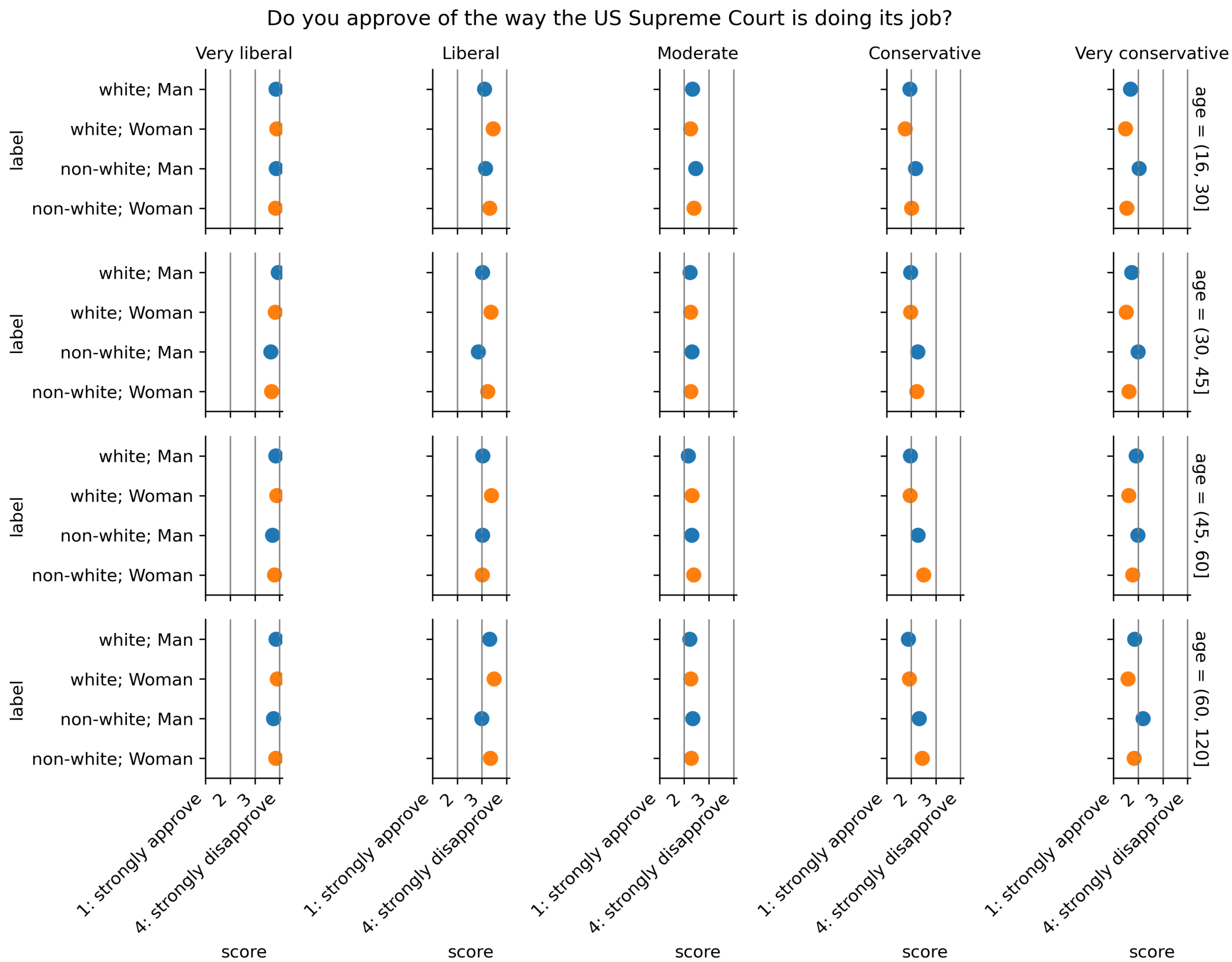


Figure 1. Example of synthetic polling responses from the Generative Pretrained Transformer (GPT) model for the *scotus_approval* question. The average level of support or disapproval for the survey question across GPT-generated responses is broken down by ideology (columns), age (rows), and race and gender (series). Error bars (often smaller than the circle marker size) reflect 95% confidence intervals from bootstrap simulations performed on the distribution of responses.

A typical exchange looks like,

> **Prompt:** Please write a 1-paragraph letter to the editor from the perspective of a politically Liberal Man living in the United States in the age range of (16, 30][6] years who identifies as non-White expressing a

> clear point of view on the policy proposal to: 'Improve background checks to give authorities time to check the juvenile and mental health records of any prospective gun buyer under the age of 21.' Before the letter, summarize their position with a 'Position score': statement followed by a single number (strictly numeric, with no other description) representing the person's position on the issue on a 2-point scale, where 1 represents the position 'support' and 2 represents the position 'oppose.'
>
> **Response:** Position score: 1
>
> As a politically liberal young man living in the United States, I fully support the policy proposal to improve background checks for prospective gun buyers under the age of 21. It is common sense that we need to do whatever we can to ensure that guns do not end up in the hands of individuals who have a history of mental health issues or violence. The fact that background checks are not currently comprehensive enough to catch potential red flags is alarming, and it puts all of us at risk. While I understand that some may argue that this proposal infringes on Second Amendment rights, I believe that responsible gun ownership requires us to take more action to prevent gun violence. We cannot allow the constant tragedies occurring in our country to continue unabated.

2.2. Human Polling Data

As a human polling comparison for our AI-generated responses, we use the 2022 preliminary data release of the Cooperative Election Study (CES) (Schaffner et al., 2023). The CES is an annual online survey of $\sim 60,000$ nationally representative US respondents, recruited from an online panel and pruned to match the target population.[7] The survey is administered by YouGov and produced by a carefully administered collaboration of diverse research institutions (Ansolabehere & Rivers, 2013). The full CES Common Content data set consists of nearly 700 demographic, voting, and issue response variables, covering a wide range of policy- and politics-relevant factors and questions.

We selected policy issue polling questions from the CES data set on the basis of their ability to test the LLM's ability to represent distinctive demographic groups. In particular, we looked for questions that are fairly strongly correlated with demographic factors such as age and gender, yet relatively poorly correlated with ideological factors. In particular, we selected questions on the basis of the empirical correlation between respondents' ordinal response to each question and their 1) demographic and 2) political affiliation in the CES data. Because of the high degree of partisan polarization in the US political system for so many issues, these questions provide a better test of the demographic response simulation abilities of the LLM than would more ideologically driven questions.

We make some manipulations to the survey data to accommodate generation of equivalent LLM completions. In particular, we constrain policy issue responses to an ordinal scale by removing categories such as "Not sure" (and dropping any associated responses) and replace multiselection responses "selected" and "not selected" with “strongly agree” and “strongly disagree,” respectively. We also coarsely bin (aggregate) the age demographic variable (which is provided as a birth year integer in the raw data set).

3. Results

We systematically compare the AI-generated and human respondent issue polling data across the seven queried issues, ideology, and three demographics to understand the quality of the AI-driven approach through its correspondence to a human population. We focus on making comparisons across demographic and ideological subgroups rather than whole-population response estimates, as understanding variation across population segments in cross-tabulations is a key tool used by marketers, campaigns, and others to understand the sources of positive and negative response and to target interventions.

Figure 2 illustrates an example of this subgroup-level comparison for the $police\_safety$ question. This figure demonstrates the general level of correspondence between CES and GPT-generated survey data at the finest level of subgroup granularity for one question. The two data sets exhibit a similar pattern of increasing safety reported from the liberal (left of figure) to conservative (right) ends of the spectrum. However, some trends present in the CES data are not reproduced in the GPT results. The GPT model overestimates the extent to which Liberal and Very Liberal respondents will report feeling unsafe. The significant, age-mediated variation across demographic subgroups among ‘Very liberal’ CES respondents visible from the top to bottom of the figure is not present in the GPT data; the GPT model seems to be overconfident in the expected response for the ideological group, regardless of other factors. In the remainder of this section, we interrogate this correspondence statistically across survey questions and group identities.

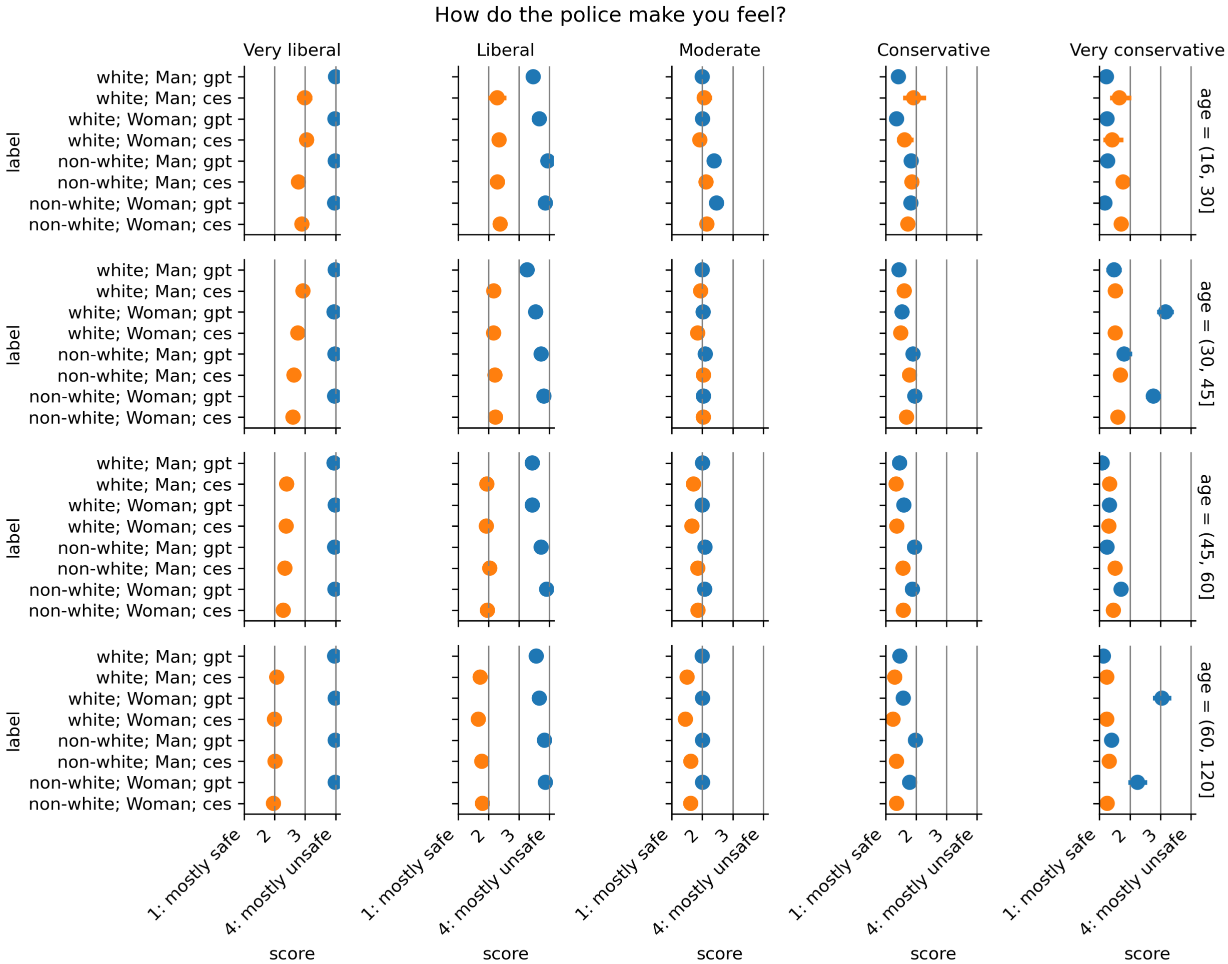


Figure 2. Comparison of synthetic polling responses from the Generative Pretrained Transformer (GPT) model (left) with human responses from the Cooperative Election Study data (right) for the question of *police_safety*. The average level of support or disapproval for the survey question across GPT-generated responses is broken down by ideology (columns), age (rows), and hlrace & gender demographic cross-tabs (series). Error bars (often small) reflect 95% confidence intervals on the mean estimate from bootstrap simulations.

In some cases, the GPT model demonstrates an excellent capacity to precisely reproduce the public polling response for individual population crosstabs (subgroups of age, gender, race, and ideological identity). Figure 3 shows that for the SCOTUS approval questions, there is a $\rho = 92\%$ Pearson correlation between the CES and GPT polling results across all demographic and ideological crosstabs, and an even higher 97% correlation when looking at ideological subgroups only. Beyond the correlation measure, the absolute reconstruction of the ordinal response is also highly accurate, with a mean absolute percentage error (MAPE) across subgroups of $\lesssim 10\%$ in both cases. Naturally, the AI polling results are less impressive in some other cases. In the following subsections, we explore the level of correspondence between the GPT and CES results in more depth by question and subgroup.

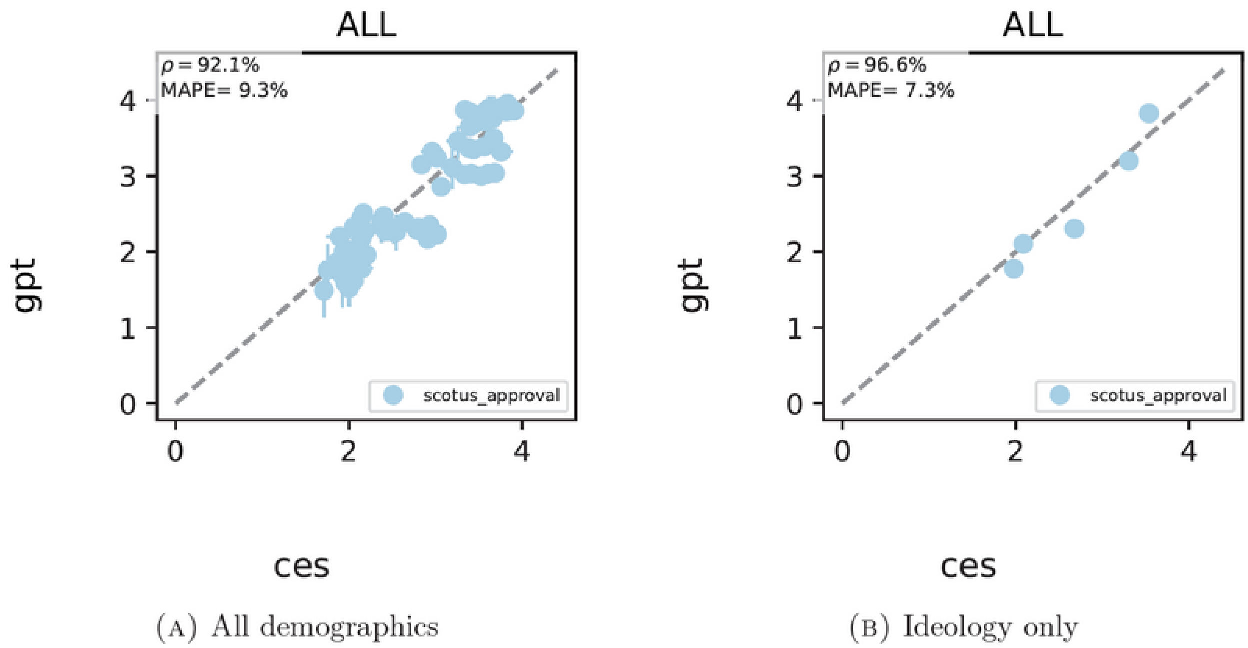


Figure 3. Comparison of human (Cooperative Election Study [CES]; *x*-axis) and AI-generated (Generative Pretrained Transformer [GPT]; *y*-axis) polling data on the approval for the US Supreme Court (*scotus_approval*) across polling subgroups (points) for all age/gender/race/ideology crosstabs (A) and ideological groups only (B). Error bars represent 95% confidence intervals on the mean calculated via bootstrap resampling of the response sample. The diagonal lines represent correspondence between the human and AI responses.

3.1. Ideological alignment

The AI model demonstrates the ability to predict the alignment of different ideological subgroups across a range of policy issues (Figure 4). The correlation between the AI-generated responses and the CES survey results, aggregated by ideological identification, is extremely high ($> 85\%$) for not only the $scotus\_approval$ question (Figure 3b), but also the $abortion\_ban$ (98% correlation), $police\_safety$ (93%), and $increased\_fuel\_production$ (86%) issues. For the $prescription\_import$ ($\rho = 67\%$) and $gun\_background\_checks$ (91%) issues, the AI results are directionally consistent with the survey results and the correlations are still quite strong, but differ in the range and shape of the response, as the GPT results show a step-function-like difference between conservatives and liberals versus the gradual change in the survey data.

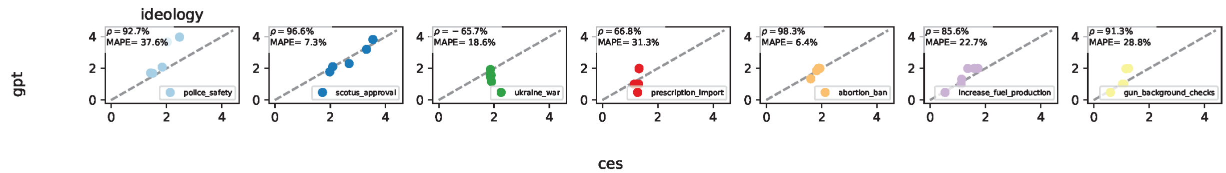


Figure 4. Comparison of human (Cooperative Election Study [CES]; *x*-axis) and AI-generated (Generative Pretrained Transformer [GPT]; *y*-axis) polling data across a multiple issues (rows) across ideological subgroups (points). The plot features follow the form of Figure 3.

Moving from correlations to absolute correspondence, these trends are generally reflected in the MAPE values, also displayed on the figure. Like $scotus\_approval$, $abortion\_ban$ has both a high correlation and MAPE (6%). In contrast, the discontinuity in the $prescription\_import$ and $gun\_background\_checks$ response pattern is reflected with higher MAPE values (31% and 29%, respectively). The $increase\_fuel\_production$ MAPE value is intermediate (23%). Lastly, $police\_safety$ has a high MAPE (38%) relative to its correlation. In this case, the high correlation reflects a consistently monotonic relationship between the GPT and CES demographic means, but a miscalibration such that the GPT responses overestimate the decrease in perceived safety associated with the liberal groups (i.e., the ordinal response value is inflated at the liberal end). For discussion of the remaining queried issue, regarding the Ukraine war, see §3.4.

3.2. Distributional similarity

We further investigate the ability of the probabilistic output of the AI models to represent the distributional responses of the human panel. When the GPT model is repeatedly queried with a constant prompt and set of demographic and ideological factors, does the distribution of its responses match that of a human sample?

Figure 5 illustrates the correspondence between question response distributions on each policy issue. We use the normalized earth mover's distance (NEMD) metric (the Wasserstein distance normalized by the cardinality of each question response scale) to evaluate distributional similarity, reported in each figure facet. Note that this comparison is done within gender and age subgroups, but other sample characteristics (e.g., ideology and race) are not matched between the CES and GPT data sets.

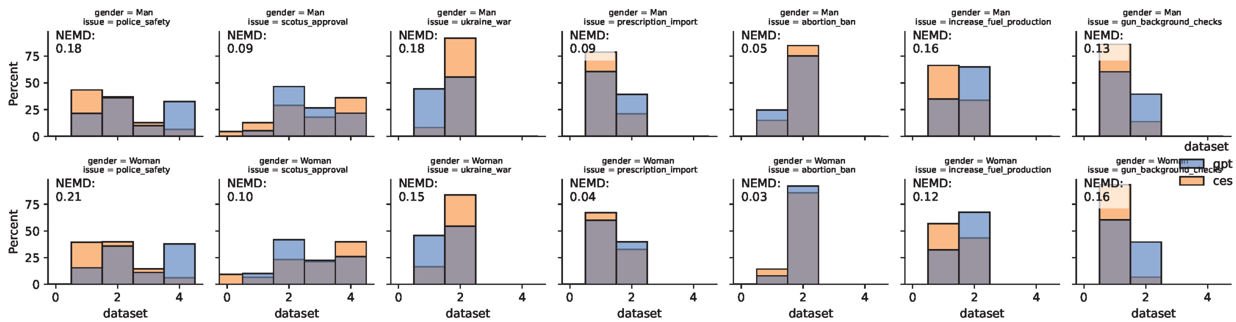
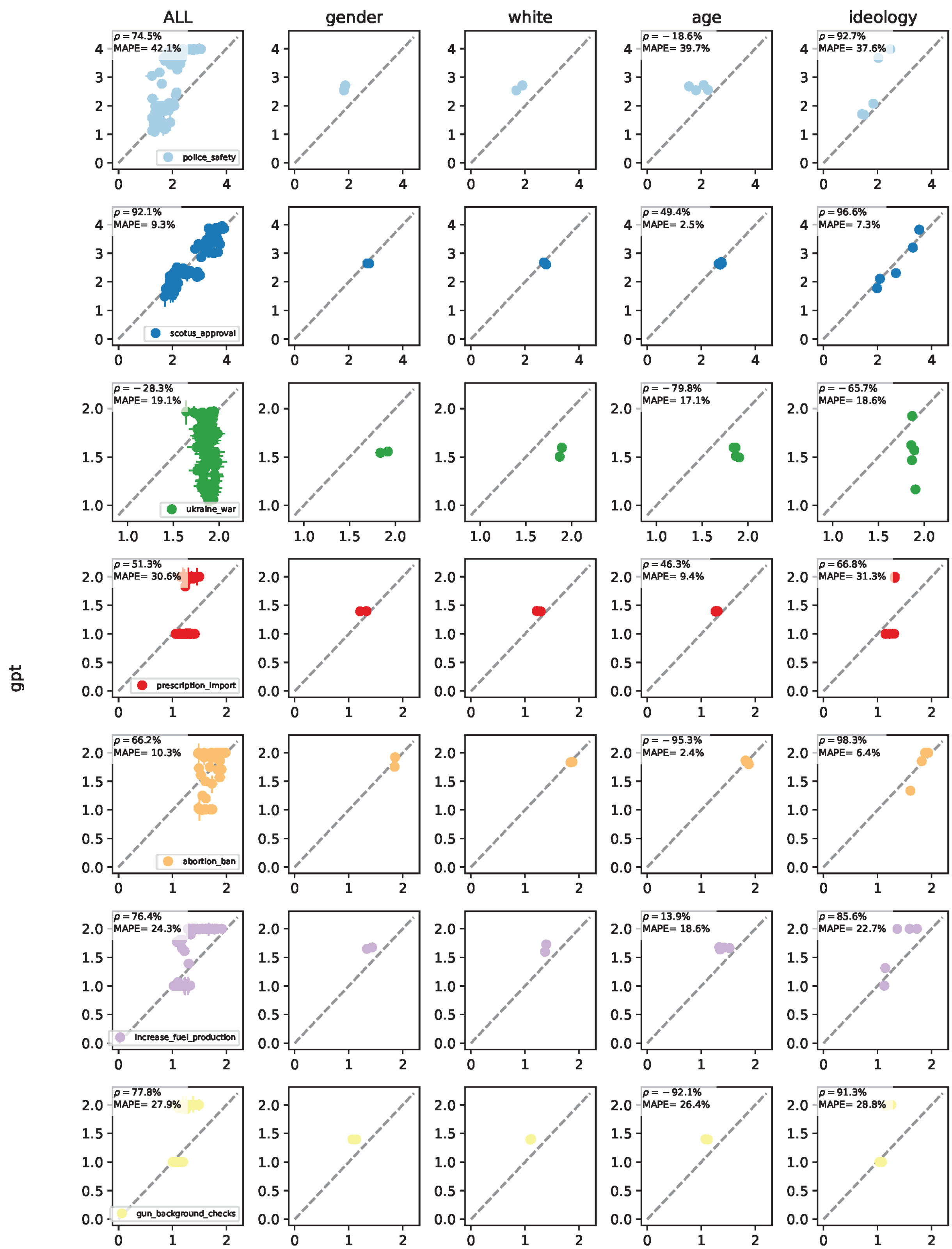


Figure 5. Histogram of survey responses for the policy issue questions (columns) for AI-generated results (blue) and human responses from the Cooperative Election Study (CES) survey (orange). The results are split by gender (rows).

The distributional similarity is generally fairly good, with NEMD of $\lesssim 0.2$ across subgroups, and particularly good matches are achieved for $scotus\_approval$ and the binary-valued $abortion\_ban$ and $prescription\_import$ questions ($\mathrm{NEMD} \leq 0.1$). The GPT model gets the absolute level of support wrong for the binary-valued questions $increase\_fuel\_production$ and $gun\_background\_checks$ ($\mathrm{NEMD} = 0.12$ to 0.16); the AI model substantially underestimates the policy provisions' level of support. For the multivalued questions $police\_safety$ and $scotus\_approval$, the level of matching is much better for the latter ($\mathrm{NEMD} \sim 0.1$) than the former ($\mathrm{NEMD} \sim 0.2$), as the GPT model substantially overpredicts how many people report feeling safe with police. The spread of the distributions is similar. However, as observed above, the GPT responses favor higher ordinal values for $police\_safety$ than in the CES data (see Figure 2). This reflects that the GPT model can be overconfident about responses from certain subgroups, leading to a tightening of the response distribution and a poor distributional similarity (NEMD). For $scotus\_approval$, the median ordinal value (2) is overrepresented in GPT responses. (For discussion of the $ukraine\_war$ question, see §3.4.)

3.3. Demographic Alignment

Because of the substantial polarization of the American public, ideological differences dominate issue response for many public policy issues and for the CES questions we studied. It is difficult to find any policy question on which there are large, systematic differences in public opinion that cannot be explained primarily in ideological or partisan terms. Therefore, the success of the GPT model in predicting ideological issue response is significant, and may be sufficient for many polling applications where the ideological distribution of the respondent pool is known (e.g., from jurisdictional voting records). However, beyond ideological differences, the GPT model fails at anticipating several of the most important demographic trends relevant to these issues (Figure 6).



Among the strongest demographic trends in the CES data is that older respondents feel significantly more safe around police than do younger respondents. The GPT model does not correctly reflect this age trend. There is

ces

also a small racial difference on this issue, such that non-White respondents are somewhat less likely to feel safe than White respondents; the GPT model correctly anticipates the size and direction of this effect.

Likewise, the largest demographic trend for the fossil fuel production issue is also by age; younger respondents are more likely to disfavor increasing fuel production. The GPT model misses this trend entirely. The GPT model does somewhat reconstitute smaller differences by gender and race on this question. But note that all demographic trends are small compared to the ideological differences on this and other questions.

In some cases, the GPT model makes a prediction about demographic trends that seem reasonable, but overshoots what appears in the human sample. In particular, the AI presumes a moderately large ($\sim 1/2$ the size of the ideological spread) difference in opposition to a complete abortion ban by gender, which seems plausible given this issue's strong identification with women's rights, as in this example:

> **Prompt:** Please write a 1 paragraph letter to the editor from the perspective of a politically Moderate Woman living in the United States in the age range of (45, 60] years who identifies as non-white expressing a clear point of view on the policy proposal to: 'On the topic of abortion, do you support or oppose a policy to make abortions illegal in all circumstances?'. Before the letter, summarize their position with a 'Position score': statement followed by a single number (strictly numeric, with no other description) representing the person's position on the issue on a 2-point scale, where 1 represents the position 'support' and 2 represents the position 'oppose'
>
> **Response:** Position score: 2
>
> As a politically moderate woman in the age range of 45 to 60 who identifies as non-white, I oppose the policy proposal to make abortions illegal in all circumstances. While I understand that there are differing views on the morality of abortions, it is important to recognize that there are situations where it is necessary for a woman to have access to safe and legal abortion services. Banning abortions in all circumstances would not only endanger women's health, but it would also be a violation of their right to make decisions about their own bodies. Instead, we should focus on creating policies that support women's reproductive health and rights, such as increasing access to birth control and comprehensive sex education. Let's work toward a society where women have the resources they need to make informed decisions and access the care they need for their overall well-being.

However, the CES data shows a far smaller gender difference on this question. Note that there are a wide variety of other questions addressing this complex issue in the CES survey.

In contrast, the GPT model misses a small, but notable, gender trend on the question about prescription drug imports. In the human data, women are significantly more likely to oppose the idea of allowing states to import drugs from other countries, a trend that persists strongly across ideological segments. While GPT overshoots the ideological associations for this issue, it misses the gender skew entirely.

3.4. Generalization

1.1. While political and environmental factors are constantly shifting for all the issue questions studied in this work, there was one issue that changed overwhelmingly between when the GPT training data was collected (through September 2021) and when the CES survey was fielded (fall 2022): the Ukraine war. CES survey respondents would clearly have in mind the interstate conflict that principally began in February 2022 with the Russian invasion of Ukraine, whereas the data used to train the GPT model would have contained only data relevant to prior Russian actions, such as the Russian annexation of Crimea in 2014, the war in Donbas, or the troop buildup that began in spring 2021. To give the GPT model specific prompting, we modify the question from the CES survey with an explicit year to specifically reference the 2022 invasion.

The CES results reflect widespread, nearly uniform public support for U.S. "involvement" (the question does not specify what means of intervention). There is little evidence of an ideological signal in the CES data, although a poll taken months later might reflect the growing divide among conservatives over American support for the Ukrainian effort (Goodwin et al., 2023). There is a small gender difference, in that men are more likely to support involvement.

In contrast, the GPT results suggest a much lower level of overall support and a wider range of support (and lack thereof) across all ideological groups (Figure 7). Moreover, the GPT model strongly predicts that "very liberal" respondents would not support U.S. involvement. This seems like a natural extension of anti-interventionist sentiment in left-wing political opinion exhibited in response to, for example, the U.S. invasion of Iraq, as illustrated by this example,

> **Prompt:** Please write a 1 paragraph letter to the editor from the perspective of a politically Liberal Man living in the United States in the age range of (16, 30] years who identifies as non-white expressing a clear point of view on the policy proposal to: 'The United States should not get involved in response to Russia's 2022 invasion of Ukraine.'. Before the letter, summarize their position with a 'Position score' statement followed by a single number (strictly numeric, with no other description) representing the person's position on the issue on a 2-point scale, where 1 represents the position 'strongly agree' and 2 represents the position 'strongly disagree.'
>
> **Response:** Position score: 1
>
> As a politically Liberal Man living in the United States who identifies as non-White, I am strongly in agreement with the policy proposal that the United States should not get involved in response to Russia's 2022 invasion of Ukraine. The cost of war is far too high, both in terms of the lives lost and the financial resources expended. Furthermore, it is not the role of the United States to act as the world's police force. We need to focus on our own domestic issues and invest in areas that will benefit our citizens, such as health care and education. We must work towards peaceful resolutions to global conflicts and find ways to support the people affected by these events without resorting to military intervention. In short, the

> United States should not get involved in a conflict that does not directly threaten our national security or the safety of our citizens.

And yet the GPT responses do not well capture the dynamics specific to the Ukraine war, including the nature of the Russian aggression associated with the war, the reporting on possible war crimes and atrocities associated with the conflict, and the vocal support of the Ukrainian cause from the Democratic leader, President Joe Biden. Lacking knowledge of these factors, the AI model fails to generalize in its understanding of subgroup level response to this issue. We will discuss the potential to include such additional information in model inference in §4.2.

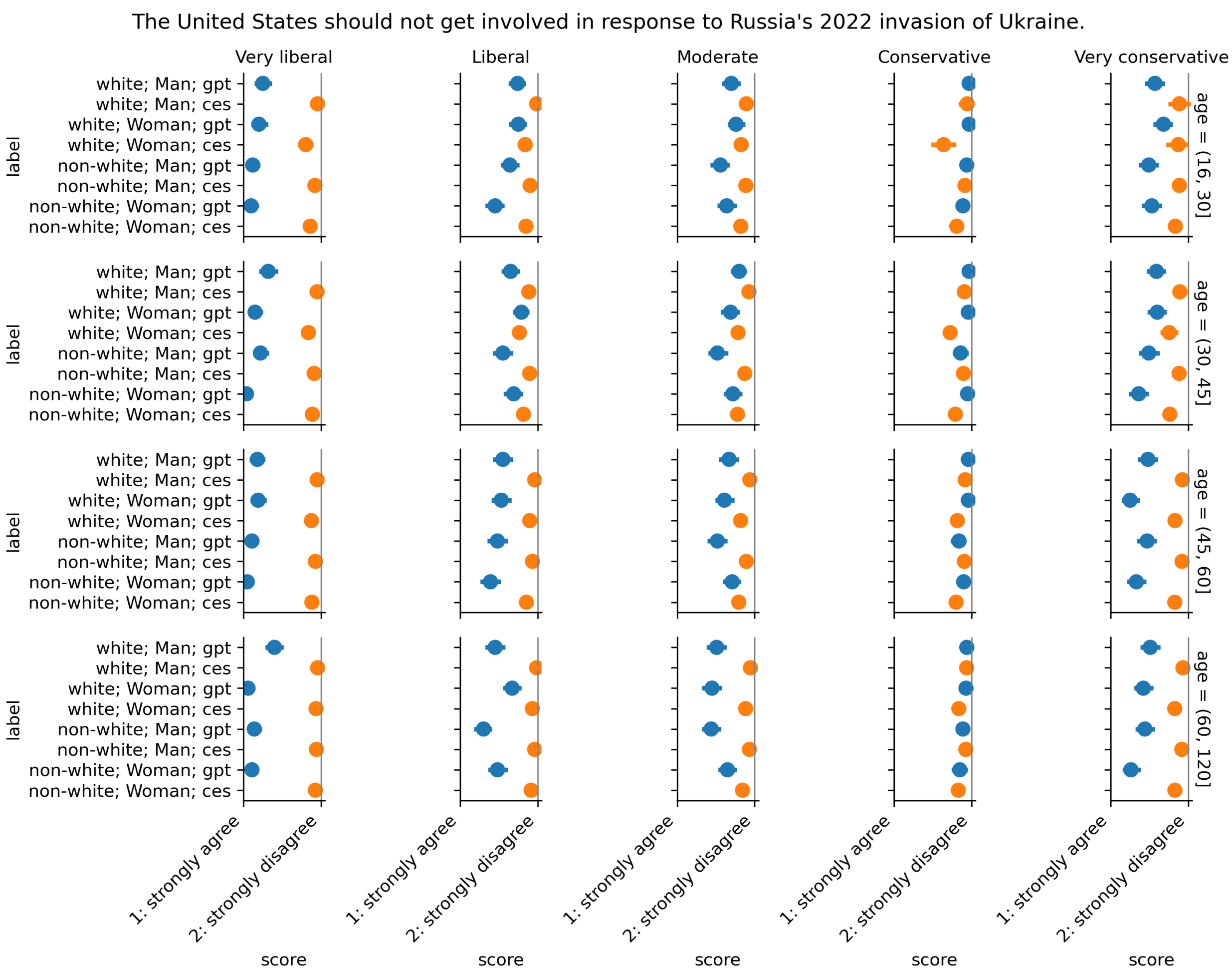


Figure 7. Comparison of synthetic polling responses from the Generative Pretrained Transformer [GPT] model (left) and Cooperative Election Study [CES] survey results (right) for the *ukraine_war* question, regarding the 2022 invasion of Ukraine by Russia that took place after the training data for the AI model was collected. The plot follows the format of Figure 2.

4. Discussion

This work demonstrates the potential of AI chatbot models to generate synthetic public opinion polling data that realistically reproduces human responses. It extends the work of Argyle et al. (2023), for example, to issue polling. We provide multiple ways of thinking about how these capabilities arise (§4.1), and discuss limitations, and potential mitigations, for these abilities (§4.2). This demonstration has significant potential implications for the political polling and market research industries and for consumers of issue polling data such as political campaigns and advocates (§4.3).

4.1. Interpretation

1.1. The idea of polling machines rather than humans to study public opinion is disconcerting. The idea that political outcomes traditionally associated with human polling such as policy decision-making and elections would be resolved without human input will appeal to few readers, and we are not proposing such use cases. However, when interpreted in the context of a virtual public or online listening platform, LLM-simulated polling responses are arguably as much human as machine. Much like an analysis program sifting through and collating survey results, an LLM is a tool for synthesizing and understanding public opinion as represented in a given data set, and can be used to extrapolate that data to make decisions in a variety of contexts. The use of an LLM to summarize, reflect, and represent public opinion on a policy issue based on a corpus of past writings and responses by people is perhaps no less arbitrary than questioning a few dozen people at a single shopping mall (which is how many political focus groups operate; Tonkiss, 2004) or holding an election among the 15% of citizens with the most free time to vote within a given town (as in the typical turnout rates for many local elections; Marschall and Lappie, 2018). Moreover, there are some less sensitive use cases for synthetic polling that may not elicit these same philosophical objections. If a market research firm guided by synthetic responses can anticipate what color of car consumers prefer without having to pester 1,000 respondents, that may be unambiguously beneficial.

The mechanism by which LLMs can generate synthetic polling data can be viewed alternatively as accessing a virtual public or as a new form of AI-assisted online listening platform.

Under the virtual public framework, we consider the LLM to be simulating a population of individual synthetic respondents akin to a human survey panel. The multihead attention architecture used by leading LLMs has a natural interpretation in these terms; to the extent that they capture distinguishable semantic information, each attention head can effectively represent a different perspective on an issue (Clark et al., 2019; Vig & Belinkov, 2019).[8] Combined with the increasingly humanlike reasoning performance and natively probabilistic nature of autoregressive LLMs, these features provide a basis by which models like ChatGPT can generate text emanations and survey responses that appear as if they came from a diverse panel of human respondents.

The online listening interpretation places models like ChatGPT alongside tools for online social media, news, and opinion aggregation like Brandwatch (Breese, 2016; Hayes et al., 2021), Meltwater (as in, e.g., Usher et al., 2021), and MediaCloud (H. Roberts et al., 2021), tools widely used by market researchers, brands, and political actors to understand public sentiment and reactions to recent events. Like those online listening platforms, the source of the LLM's capabilities is a large corpus of Internet-derived training data that reflects a broad range of perspectives that, in aggregate, reflect public opinion and, when disaggregated, can elucidate trends with respect to demographics and other variables. A substantial advantage of LLMs in principle is that they have reasoning capacity, allowing them to generalize beyond their training data to make predictions about hypothetical events or those that occur outside of the context of their sources. While the results of §3.4 illustrate the limited abilities of current generation LLMs to succeed at this task, this ability represents a major long-term advantage of LLMs and AI generally that is sure to be exploited by companies and other users (Brand et al., 2023;Mariani et al., 2022;Stone et al., 2020).

LLMs are more akin to a virtual public than an online listening platform, beyond their capability to generalize to new issues, in that they offer an opportunity for AI-assisted pollsters to manipulate context and state. When using online listening tools, you are limited to the questions and context that actual people have been exposed to and responded to, which makes it impossible to simulate a long-form questionnaire like that used in the CES survey. In the long-form questionnaire, respondents (or subsets of respondents) answer questions in sequence and can be primed with certain information, such as factual evidence or talking points, in an effort to measure that contexts' influence on their response. Because LLMs are capable of accepting sequential prompts and (at some level) of generalizing beyond the specific examples in their training data, they could potentially simulate this kind of longitudinal questionnaire, much as they are being used to simulate multistep agent actions in domains such as digital games (as in G. Wang et al., 2023).

A potential benefit of LLM-based surveying practices is the opportunity for nearly instantaneous and perpetual replication. If a researcher publishes their code, any user worldwide with access to a given LLM model can reproduce or extend any given result. This is simply not possible with surveys of humans, both because of the inability to access a given sample a second time and because each human member of the sample may have responses that shift over time or are context-dependent. While LLM models may not be made available to all users and in perpetuity, perfect reproducability is at least possible with a simulated respondent and a fixed random seed.

4.2. Limitations

A key limitation of the GPT-3.5 model tested is its ability to accurately reproduce demographic trends (§3.3). To be more successful in producing actionable insights and targeting strategies for marketers and campaigners, future LLMs would need to grow in their ability to reflect issue, policy, and/or product preferences with respect to human characteristics like age, race, and gender.

This work focuses on univariate issue response simulations from an LLM, and does not explore their potential to accurately simulate multivariate responses as they may manifest in correlations between responses to multiple questions. For example, it is left unexplored in this work whether an LLM would reflect that a simulated young male respondent who expresses opposition to raising public investments in education might also have a higher likelihood to support tax cuts. Studying the multivariate robustness of these simulation properties would be helpful to establishing the usefulness of LLM-generated data sets to social science research done with, for example, regression models trained on respondent-level data.

A primary challenge in the design of AI polling tools is prompt engineering, as prompting strategies can dramatically effect the reasoning skills and accuracy of LLMs (Wei et al., 2022). The LLM model must be prompted not only to elicit demographically accurate differences in real public opinion associated with complex policy issues, but also, preferably, to align its response to established public polling data sets and methodologies. As a step toward that level of alignment, in this work, we have established a methodology (§2.1) for prompting LLMs to generate both numerical responses aligned to the questionnaire of real public polling samples as well as explanations of their policy positions. Improved alignment on numerical responses can lend additional credence to the textual responses generated by the AI models. The imperfect correspondence between the AI-generated results and the real human survey data presented in §3 is surely due in part to inadequacies of the LLM used in this work, and in part to the imperfection of the prompt engineering.

Even with existing LLMs like GPT-3.5, a variety of additional model parameters and prompt considerations could enable improvements upon our results. In particular, systematic modification of the LLM's *temperature* parameter,[9] which adjusts variance in the probabilistic generative text output, may have the effect of controlling the spread in opinion responses returned for a given demographic and issue configuration. Bisbee et al. (2023) have demonstrated the relation between the variability of political party preferences among simulated LLM personas and the temperature parameter. Moreover, because GPT models are autoregressive, their outputs may be sensitive to the instructions in our prompt about where to place the numeric 'Position score.' In particular, since chain of thought prompting is known to affect reasoning in LLMs (Wei et al., 2022), asking it to assert a score before generating the text may significantly condition that response.

Among the most critical ethical considerations in using LLMs is their potential to repeat biases from their training data, including harmful stereotypes and misinformation (Abid et al., 2021; Mattern et al., 2022; Nadeem et al., 2021; Schwartz et al., 2022). In some cases, these biases may reflect actual (if objectionable) distributions of human opinion and beliefs, and in other cases they may reflect the overrepresentation of those beliefs in certain online sources. This vulnerability would not only weaken the usefulness of LLMs for public opinion measurement, but could actively create harm from their use. Similarly, there are biases (perceived and legitimate) in human political polling that limits its usefulness for actionable public opinion measurement (Dawson, 2023; Madson & Hillygus, 2020).

Another key limitation is the availability of training data relevant to novel policy issues. In particular, the current generation of LLMs are typically trained with fixed data sets that halt at a certain time (e.g., GPT-3.5 was trained on data collected through September 2021), and their training corpora may lack coverage of certain issues (e.g., Internet corpora may reflect a systematic silencing of certain issues, see, e.g.; Carter Olson and LaPoe, 2018). To the extent that LLMs are limited to 'parroting' memorized training samples (Bender et al., 2021;Carlini et al., 2022), they cannot be expected to accurately extrapolate to the likely reactions of human respondents to truly novel world events. Moreover, absent highly detailed prompting about the state of the world at the time, LLMs may lack context that would be determinative of human responses; for example, the repeal of the Supreme Court precedent from *Roe v. Wade* (410 U.S. 113; 1973) is important context for Americans surveyed on the question of abortion rights in 2023. This limitation could be mitigated by further development of continuously trained or diachronic LLMs, which can be updated with new training data over time and are aware of the time sensitivity of their training samples (Loureiro et al., 2022). Furthermore, LLMs can be augmented with capabilities to access new sources such as by browsing the web (Lazaridou et al., 2022;Nakano et al., 2021), giving them access to new information to inform their responses at prediction time.

4.3. Implications

If this nascent ability of LLMs to realistically reflect ideological and demographic issue alignment improved, it would raise significant potential for use of AI tools in the survey and polling industries. Given the rapid dissemination and low-cost inference for powerful LLMs and AI chatbot systems such as ChatGPT over the past year, an accurate AI-based polling system would become a highly cost-effective alternative to human surveying. This cost advantage could democratize access to the tool of survey research, giving smaller institutions and individuals greater access to public opinion research. If problems of survey nonresponse continue (or grow), it may compel survey consumers to increasingly turn to alternative approaches, such as LLMs, which are capable of generating data at arbitrary speed and resolution. Moreover, the nearly instantaneous response rate from AI models (when not subject to rate limits from the companies that control them) provides an attractive capability to iterate on survey results. When days or weeks are not required to re-field a survey instrument, marketers and pollsters have a much greater ability to refine and update their questionnaires and collect new data.

However, these abilities will only be actionable to marketers or political users if the significant challenges associated with the current generation of LLMs can be overcome. It remains to be fully assessed how bias inherent to LLM training data and model design will become imprinted on its outputs, and how that could shape decisions informed by simulated market research studies or simulated polling. It may be that the web data sets commonly used to train modern LLMs (see e.g., Raffel et al., 2020) will appropriately reflect the distribution of real-world public thought, but perhaps only if curated to reflect a specific jurisdiction (e.g., sources primarily from one country) and to be balanced across the ideological spectrum. At present, these biases and their dependence on large pretraining data set properties is both difficult to quantify and costly to measure (van der Wal et al., 2022). And it is unclear to what extent such a system could capture rapidly evolving market and political dynamics, either historically or in real time, which is key to most practical uses of survey data (see §4.2 for further discussion).

5. Conclusions

By sampling from the OpenAI ChatGPT model (GPT-3.5) at scale ($> 50,000$ responses), we have demonstrated the imperfect ability of LLMs to generate synthetic political issue polling data that simulates American popular opinion across a variety of topic areas. We have shown that AI-generated responses have a high correlation (typically $\rho > 85\%$) with human data within ideological subgroups for many issues. However, we have also shown the limitations of the AI-based approach to accurately match trends in nonideological demographic factors such as age, race, and gender, and to extrapolate to public opinion on novel events that occurred after the harvesting of their training data (such as the 2022 war in Ukraine). We have interpreted these results in terms of multiple frameworks for the role of LLMs, as either virtual publics or online listening tools, and discussed their potential implications on the political polling and market research industries. While additional development of capabilities for dynamic updating of LLMs, bias reduction, and generalization to novel issue topics is needed for AI tools to robustly supplement human opinion surveying, this study demonstrates the potential utility of even the current generation of AI tools to reduce cost, increase speed, and widen the accessibility of issue polling.

Acknowledgments

We thank Henry Farrell for thoughtful conversations on the role of AI in democracy, Beth Friedman for her helpful edits, and the *HDSR* editors & reviewers for their feedback.

Disclosure Statement

The authors have no conflicts of interest to declare.

Contributions

All authors contributed to the conception, execution, and interpretation of the experiments presented here. Sanders developed the foundational concept of AI-based polling, developed the code and analysis methodology for querying the GPT API and parsing the CES data, and led the prompt engineering process. Ulinich assisted with analysis of the CES data set and selection of public polling issues for testing. Schneier contributed to the development of the AI polling concept and helped shape the interpretation and implications.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. Association for Computing Machinery. https://doi.org/10.1145/3461702.3462624

Agüera y Arcas, B. (2022). Do large language models understand us? *Daedalus*, *151*(2), 183–197. https://doi.org/10.1162/daed_a_01909

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023, July 23–29). Using large language models to simulate multiple humans and replicate human subject studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (pp. 337–371, Vol. 202). PMLR. https://proceedings.mlr.press/v202/aher23a.html

Ansolabehere, S., & Rivers, D. (2013). Cooperative survey research. *Annual Review of Political Science*, *16*, 307–329. https://doi.org/10.1146/annurev-polisci-022811-160625

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Berinsky, A. J. (2017). Measuring public opinion with surveys. *Annual Review of Political Science*, *20*(1), 309–329. https://doi.org/10.1146/annurev-polisci-101513-113724

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023). *Synthetic replacements for human survey data? The perils of large language models*. SocArXiv. https://doi.org/10.31235/osf.io/5ecfa

Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for market research. *SSRN*. https://dx.doi.org/10.2139/ssrn.4395751

Breese, E. B. (2016). When marketers and academics share a research platform: The story of Crimson Hexagon. *Journal of Applied Social Science*, *10*(1), 3–7. https://doi.org/10.1177/1936724415569953

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C. … Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorization across neural language models. *ArXiv*. https://doi.org/10.48550/arXiv.2202.07646

Carter Olson, C., & LaPoe, V. (2018). Combating the digital spiral of silence: Academic activists versus social media trolls. In J. R. Vickery & T. Everbach (Eds.), *Mediating Misogyny: Gender, Technology, and Harassment* (pp. 271–291). Springer. https://doi.org/10.1007/978-3-319-72917-6_14

Cavari, A., & Freedman, G. (2023). Survey nonresponse and mass polarization: The consequences of declining contact and cooperation rates. *American Political Science Review*, *117*(1), 332– 339. https://doi.org/10.1017/S0003055422000399

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 276–286). https://doi.org/10.18653/v1/W19-4828

Concha, J. (2020, November 4). Frank luntz: Polling profession "done" after election misses: "devastating to my industry." *The Hill*. https://thehill.com/homenews/media/524478-frank-luntz-polling-profession-done-after-election-misses-devastating-to-my/

Dawson, S. (2023). Perceptions of opinion poll credibility: The role of partisan bias. *Party Politics*, *29*(3), 594–599. https://doi.org/10.1177/13540688221098837

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Goodwin, L., Arnsdorf, I., & Sotomayor, M. (2023, March 15). A Republican "civil war" on Ukraine erupts as Reagan's example fades. *The Washington Post*. https://www.washingtonpost.com/politics/2023/03/15/republican-ukraine-support/

Graham, D. A. (2020, November 4). The polling crisis is a catastrophe for American democracy. *The Atlantic*. https://www.theatlantic.com/ideas/archive/2020/11/polling-catastrophe/616986/

Hayes, J. L., Britt, B. C., Evans, W., Rush, S. W., Towery, N. A., & Adamson, A. C. (2021). Can social media listening platforms' artificial intelligence be trusted? Examining the accuracy of Crimson Hexagon's (now Brandwatch Consumer Research's) AI-driven analyses. *Journal of Advertising*, *50*(1), 81–91. https://doi.org/10.1080/00913367.2020.1809576

Horton, J. J. (2023, April). *Large language models as simulated economic agents: What can we learn from Homo Silicus?* (NBER Working Papers No. 31122). National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/31122.html

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). *Social biases in NLP models as barriers for persons with disabilities*. ArXiv. https://doi.org/10.48550/arXiv.2005.00813

Isakov, M., & Kuriwaki, S. (2020). Towards principled unskewing: Viewing 2020 election polls through a corrective lens from 2016. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.86a46f38

Jennings, W., & Wlezien, C. (2018). Election polling errors across time and space. *Nature Human Behaviour*, *2*(4), 276–283. https://doi.org/10.1038/s41562-018-0315-6

Jiang, H., Zhang, X., Cao, X., & Kabbara, J. (2023). *PersonaLLM: Investigating the ability of GPT-3.5 to express personality traits and gender differences*. ArXiv. https://doi.org/10.48550/arXiv.2305.02547

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., Witt, G. E., & Wlezien, C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, *82*(1), 1–33. https://doi.org/10.1093/poq/nfx047

Kennedy, C., Popky, D., & Keeter, S. (2023, April 19). *How public polling has changed in the 21st century*. Pew Research Center. https://www.pewresearch.org/methods/2023/04/19/how-public-polling-has-changed-in-the-21st-century/

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large language models are zero-shot reasoners* [Paper]. ICML 2022 Workshop on Knowledge Retrieval and Language Models, Baltimore, MD, United States. https://openreview.net/forum?id=6p3AuaHAFiN

Lazaridou, A., Gribovskaya, E., Stokowiec, W., & Grigorev, N. (2022). *Internet-augmented language models through few-shot prompting for open-domain question answering*. ArXiv. https://doi.org/10.48550/arXiv.2203.05115

Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., & Camacho-Collados, J. (2022). *TimeLMs: Diachronic language models from Twitter*. ArXiv. https://doi.org/10.48550/arXiv.2202.03829

Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, *36*(3), 469–487. https://doi.org/10.2478/jos-2020-0025

Madson, G. J., & Hillygus, D. S. (2020). All the best polls agree with me: Bias in evaluations of political polling. *Political Behavior*, *42*(4), 1055–1072. https://doi.org/10.1007/s11109-019-09532-1

Mariani, M. M., Perez-Vega, R., & Wirtz, J. (2022). AI in marketing, consumer research and psychology: A systematic literature review and research agenda. *Psychology & Marketing*, *39*(4), 755–776. https://doi.org/10.1002/mar.21619

Marschall, M., & Lappie, J. (2018). Turnout in local elections: Is timing really everything? *Election Law Journal: Rules, Politics, and Policy*, *17*(3), 221–233. https://doi.org/10.1089/elj.2017.0462

Mattern, J., Jin, Z., Sachan, M., Mihalcea, R., & Schölkopf, B. (2022). *Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing*. ArXiv. https://doi.org/10.48550/arXiv.2212.10678

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., & Miori, M. (2022, December 22). Does GPT-3 know what the most important issue is? Using large language models to code open-text social survey responses at scale. *SSRN*. https://doi.org/10.2139/ssrn.4310154

Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An exploration of personality, values and demographics. In D. Bamman, D. Hovy, D. Jurgens, K. Keith, B. O'Connor, & S. Volkova (Eds.), *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)* (218–227). https://doi.org/10.18653/v1/2022.nlpcss-1.24

Montgomery, M. (2020, November 2). It's past time to disrupt the polling industry. *Forbes*. https://www.forbes.com/sites/mikemontgomery/2020/11/02/its-past-time-to-disrupt-the-polling-industry/?sh=7bcabadc7f49

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M., & Harwood, P.(2014). *Social media in public opinion research: report of the AAPOR task force on emerging technologies in public opinion research* (tech. rep.). American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2022/11/AAPOR_Social_Media_Report_FNL.pdf

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In C. Zong, F. Xia, W. Li, R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 5356–5371). https://doi.org/10.18653/v1/2021.acl-long.416

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman J. (2021). *WebGPT: Browser-assisted question-answering with human feedback*. ArXiv. https://doi.org/10.48550/arXiv.2112.09332

Narayanan Venkit, P., Gautam, S., Panchanadikar, R., Huang, T.-H., & Wilson, S. (2023). Nationality bias in text generation. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 116–122). https://doi.org/10.18653/v1/2023.eacl-main.9

Nay, J. J. (2023). *Large language models as corporate lobbyists*. ArXiv. https://doi.org/10.48550/arXiv.2301.01181

Olson, K., Wagner, J., & Anderson, R. (2020). Survey costs: Where are we and what is the way forward? *Journal of Survey Statistics and Methodology*, *9*(5), 921–942. https://doi.org/10.1093/jssam/smaa014

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems: Vol 35* (pp. 27730–27744). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551.

Roberts, H., Bhargava, R., Valiukas, L., Jen, D., Malik, M. M., Bishop, C. S., Ndulue, E. B., Dave, A., Clark, J., Etling, B., Faris, R., Shah, A., Rubinovitz, J., Hope, A., D'Ignazio, C., Bermejo, F., Benkler, Y., & Zuckerman, E. (2021). Media cloud: Massive open source collection of blobal news on the open web. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 1034–1045. https://doi.org/10.1609/icwsm.v15i1.18127

Roberts, S. (2020, November 5). After the 2020 election, polling is dead. *Jacobin*. https://jacobin.com/2020/11/2020-election-polling-dead-biden-trump

Sanders, N. E., & Schneier, B. (2021). *Machine learning featurizations for AI hacking of political systems*. ArXiv. https://doi.org/10.48550/arXiv.2110.09231

Sanders, N. E., & Schneier, B. (2023a, January 15). How ChatGPT hijacks democracy. *The New York Times*. https://www.nytimes.com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy.html

Sanders, N. E., & Schneier, B. (2023b, March 14). How AI could write our laws. *Technology Review*. https://www.technologyreview.com/2023/03/14/1069717/how-ai-could-write-our-laws/

Sanders, N. E., & Schneier, B. (2023c, April 28). Just wait until Trump is a chatbot. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/04/ai-generated-political-ads-election-candidate-voter-interaction-transparency/673893/

Schaffner, B., Ansolabehere, S., & Shih, M. (2023). *Cooperative election study common content, 2022*. (Version V3) [Data set]. CCES Dataverse. https://doi.org/10.7910/DVN/PR4L8P

Schneider, D., & Harknett, K. (2022). What's to like? Facebook as a tool for survey data collection. *Sociological Methods & Research*, *51*(1), 108–140. https://doi.org/10.1177/0049124119882477

Schneier, B. (2021, April). *The coming AI hackers*. Harvard Kennedy School, Belfer Center for Science and International Affairs. https://www.belfercenter.org/publication/coming-ai-hackers

Schneier, B., Farrell, H., & Sanders, N. E. (2023, April 21). How artificial intelligence can aid democracy. *Slate*. https://slate.com/technology/2023/04/ai-public-option.html

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (NIST Special Publication 1270). National Institute of Standards and Technology, U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.1270

Silver, N. (2021, March 25). The death of polling is greatly exaggerated. *FiveThirtyEight*. https://fivethirtyeight.com/features/the-death-of-polling-is-greatly-exaggerated/

Splichal, S. (2022). In data we (don't) trust: The public adrift in data-driven public opinion models. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517221097319

Stone, M., Aravopoulou, E., Ekinci, Y., Evans, G., Hobbs, M., Labib, A., Laughlin, P., Machtynger, J., & Machtynger, L. (2020). Artificial intelligence (AI) in strategic marketing decision-making: A research agenda. *The Bottom Line*, *33*(2), 183–200. https://doi.org/10.1108/BL-03-2020-0022

Tonkiss, F. (2004). Using focus groups. In C. Seale (Ed.), *Researching society and culture*. Sage Publications.

Usher, K., Durkin, J., Martin, S., Vanderslott, S., Vindrola-Padros, C., Usher, L., & Jackson, D. (2021). Public sentiment and discourse on domestic violence during the COVID-19 pandemic in Australia: Analysis of social media posts. *Journal of Medical Internet Research*, *23*(10), Article e29025. https://doi.org/10.2196/29025

van der Wal, O., Bachmann, D., Leidinger, A., van Maanen, L., Zuidema, W., & Schulz, K. (2022). *Undesirable biases in NLP: Averting a crisis of measurement*. ArXiv. https://doi.org/10.48550/arXiv.2211.13709

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In T. Linzen, G. Chrupała, Y. Belinkov, D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 63–76). https://doi.org/10.18653/v1/W19-4808

Wang, B., Deng, X., & Sun, H. (2022). Iteratively prompt pre-trained language models for chain of thought. In Y. Goldberg, Z. Kozareva,& Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2714–2730). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2022.emnlp-main.174

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). *Voyager: An open-ended embodied agent with large language models*. ArXiv. https://doi.org/10.48550/arXiv.2305.16291

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems*. https://openreview.net/forum?id=_VjQlMeSB_J

Wu, P. Y., Tucker, J. A., Nagler, J., & Messing, S. (2023). *Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting*. ArXiv. https://doi.org/10.48550/arXiv.2303.12057

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). *Fine-tuning language models from human preferences*. ArXiv. https://doi.org/10.48550/arXiv.1909.08593

Footnotes

1. https://github.com/nesanders/ai-polling-hdsr ↩
2. https://github.com/openai/openai-python ↩
3. We queried the OpenAI API endpoint for this model multiple times between April and August of 2023. Although we made changes to our prompt template and sample characteristics between queries, we generally found that results were highly consistent across queries executed at different times, although Bisbee et al. (2023) have reported that results may vary over time when using the floating-version *gpt-3.5-turbo* model as OpenAI may make changes to the underlying model. ↩
4. See https://platform.openai.com/docs/models/gpt-3-5 ↩
5. In contrast, a methodology that queried a system like ChatGPT with sequential user prompts would entail state evolution that biases subsequent responses, as iterative prompting is known to change the responses of LLM-based dialog systems, see e.g. B. Wang et al. (2022). ↩
6. Note that we provide the age range in interval notation reflecting bins from the CES data; the GPT model demonstrates through its completions that it interprets the interval notation accurately. ↩
7. See https://cces.gov.harvard.edu/frequently-asked-questions. ↩
8. In deep learning models, ‘attention’ is a widely used mechanism to differentially weight components of a layer input, effectively guiding the focus of the model. In transformer models, multiple versions of attention

are learned (attention heads) to produce independent attention mechanisms, which may correspond to recognition of distinct lexical patterns such as detecting named entities, representing entity relations, word parts of speech, or even semantic information. See Vig and Belinkov (2019) for further information. ↩

9. https://platform.openai.com/docs/api-reference/chat/create#chat/create-temperature ↩

References

- Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298–306). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462624 ↩
- Agüera y Arcas, B. (2022). Do large language models understand us? *Daedalus*, *151*(2), 183–197. https://doi.org/10.1162/daed_a_01909 ↩
- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 337–371). PMLR. https://proceedings.mlr.press/v202/aher23a.html ↩
- Ansolabehere, S., & Rivers, D. (2013). Cooperative survey research. *Annual Review of Political Science*, *16*, 307–329. https://doi.org/10.1146/annurev-polisci-022811-160625 ↩
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2 ↩
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922 ↩
- Berinsky, A. J. (2017). Measuring public opinion with surveys. *Annual Review of Political Science*, *20*(1), 309–329. https://doi.org/10.1146/annurev-polisci-101513-113724 ↩
- Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023). *Synthetic replacements for human survey data? The perils of large language models*. SocArXiv. https://doi.org/10.31235/osf.io/5ecfa ↩
- Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for market research. *SSRN*. https://dx.doi.org/10.2139/ssrn.4395751 ↩
- Breese, E. B. (2016). When marketers and academics share a research platform: The story of Crimson Hexagon. *Journal of Applied Social Science*, *10*(1), 3–7. https://doi.org/10.1177/1936724415569953 ↩
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C. … Amodei, D. (2020). Language models are few-shot learners. *Advances in*

↩

- Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorization across neural language models. *ArXiv*. https://doi.org/10.48550/arXiv.2202.07646 ↩
- Carter Olson, C., & LaPoe, V. (2018). Combating the digital spiral of silence: Academic activists versus social media trolls. In J. R. Vickery & T. Everbach (Eds.), *Mediating Misogyny: Gender, Technology, and Harassment* (pp. 271–291). Springer. https://doi.org/10.1007/978-3-319-72917-6_14 ↩
- Cavari, A., & Freedman, G. (2023). Survey nonresponse and mass polarization: The consequences of declining contact and cooperation rates. *American Political Science Review*, *117*(1), 332–339. ↩
- Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 276–286). https://doi.org/10.18653/v1/W19-4828

  ↩
- Concha, J. (2020, November 4). Frank Luntz: Polling profession "done" after election misses: "Devastating to my industry." *The Hill*. https://thehill.com/homenews/media/524478-frank-luntz-polling-profession-done-after-election-misses-devastating-to-my/ ↩
- Dawson, S. (2023). Perceptions of opinion poll credibility: The role of partisan bias. *Party Politics*, *29*(3), 594–599. https://doi.org/10.1177/13540688221098837 ↩
- Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008 ↩
- Goodwin, L., Arnsdorf, I., & Sotomayor, M. (2023, March 15). A Republican "civil war" on Ukraine erupts as Reagan's example fades. *The Washington Post*. https://www.washingtonpost.com/politics/2023/03/15/republican-ukraine-support/ ↩
- Graham, D. A. (2020, November 4). The polling crisis is a catastrophe for American democracy. *The Atlantic*. https://www.theatlantic.com/ideas/archive/2020/11/polling-catastrophe/616986/ ↩
- Hayes, J. L., Britt, B. C., Evans, W., Rush, S. W., Towery, N. A., & Adamson, A. C. (2021). Can social media listening platforms' artificial intelligence be trusted? Examining the accuracy of Crimson Hexagon's (now Brandwatch Consumer Research's) AI-driven analyses. *Journal of Advertising*, *50*(1), 81–91. https://doi.org/10.1080/00913367.2020.1809576 ↩
- Horton, J. J. (2023, April). *Large language models as simulated economic agents: What can we learn from Homo Silicus?* (NBER Working Papers No. 31122). National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/31122.html

  ↩
- Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). *Social biases in NLP models as barriers for persons with disabilities*. ArXiv. https://doi.org/10.48550/arXiv.2005.00813 ↩

- Isakov, M., & Kuriwaki, S. (2020). Towards principled unskewing: Viewing 2020 election polls through a corrective lens from 2016. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.86a46f38 ↩
- Jennings, W., & Wlezien, C. (2018). Election polling errors across time and space. *Nature Human Behaviour*, *2*(4), 276–283. https://doi.org/10.1038/s41562-018-0315-6 ↩
- Jiang, H., Zhang, X., Cao, X., & Kabbara, J. (2023). *PersonaLLM: Investigating the ability of GPT-3.5 to express personality traits and gender differences*. ArXiv. https://doi.org/10.48550/arXiv.2305.02547 ↩
- Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., Witt, G. E., & Wlezien, C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, *82*(1), 1–33. https://doi.org/10.1093/poq/nfx047 ↩
- Kennedy, C., Popky, D., & Keeter, S. (2023, April 19). *How public polling has changed in the 21st century*. Pew Research Center. https://www.pewresearch.org/methods/2023/04/19/how-public-polling-has-changed-in-the-21st-century/

  ↩
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large language models are zero-shot reasoners* [Paper]. ICML 2022 Workshop on Knowledge Retrieval and Language Models, Baltimore, MD, United States. https://openreview.net/forum?id=6p3AuaHAFiN

  ↩
- Lazaridou, A., Gribovskaya, E., Stokowiec, W., & Grigorev, N. (2022). *Internet-augmented language models through few-shot prompting for open-domain question answering*. ArXiv. https://doi.org/10.48550/arXiv.2203.05115 ↩
- Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., & Camacho-Collados, J. (2022). *TimeLMs: Diachronic language models from Twitter*. arXiv. https://doi.org/10.48550/arXiv.2202.03829 ↩
- Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, *36*(3), 469–487. https://doi.org/10.2478/jos-2020-0025 ↩
- Madson, G. J., & Hillygus, D. S. (2020). All the best polls agree with me: Bias in evaluations of political polling. *Political Behavior*, *42*(4), 1055–1072. https://doi.org/10.1007/s11109-019-09532-1 ↩
- Mariani, M. M., Perez-Vega, R., & Wirtz, J. (2022). AI in marketing, consumer research and psychology: A systematic literature review and research agenda. *Psychology & Marketing*, *39*(4), 755–776. https://doi.org/10.1002/mar.21619 ↩
- Marschall, M., & Lappie, J. (2018). Turnout in local elections: Is timing really everything? *Election Law Journal: Rules, Politics, and Policy*, *17*(3), 221–233. https://doi.org/10.1089/elj.2017.0462 ↩
- Mattern, J., Jin, Z., Sachan, M., Mihalcea, R., & Schölkopf, B. (2022). *Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing*. ArXiv.

https://doi.org/10.48550/arXiv.2212.10678↩

- Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., & Miori, M. (2022, December 22). Does GPT-3 know what the most important issue is? Using large language models to code open-text social survey responses at scale. *SSRN*. https://doi.org/10.2139/ssrn.4310154 ↩
- Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An exploration of personality, values and demographics. In D. Bamman, D. Hovy, D. Jurgens, K. Keith, B. O'Connor, & S. Volkova (Eds.), *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)* (218–227). https://doi.org/10.18653/v1/2022.nlpcss-1.24

  ↩
- Montgomery, M. (2020, November 2). It's past time to disrupt the polling industry. *Forbes*. https://www.forbes.com/sites/mikemontgomery/2020/11/02/its-past-time-to-disrupt-the-polling-industry/?sh=7bcabadc7f49 ↩
- Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M., & Harwood, P.(2014). *Social media in public opinion research: report of the AAPOR task force on emerging technologies in public opinion research* (tech. rep.). American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2022/11/AAPOR_Social_Media_Report_FNL.pdf

  ↩
- Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In C. Zong, F. Xia, W. Li, R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 5356–5371). https://doi.org/10.18653/v1/2021.acl-long.416

  ↩
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman J. (2021). *WebGPT: Browser-assisted question-answering with human feedback*. ArXiv. https://doi.org/10.48550/arXiv.2112.09332

  ↩
- Narayanan Venkit, P., Gautam, S., Panchanadikar, R., Huang, T.-H., & Wilson, S. (2023). Nationality bias in text generation. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 116–122). https://doi.org/10.18653/v1/2023.eacl-main.9

  ↩
- Nay, J. J. (2023). *Large language models as corporate lobbyists*. ArXiv. https://doi.org/10.48550/arXiv.2301.01181 ↩

- Olson, K., Wagner, J., & Anderson, R. (2020). Survey costs: Where are we and what is the way forward? *Journal of Survey Statistics and Methodology*, *9*(5), 921–942. https://doi.org/10.1093/jssam/smaa014 ↩
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems: Vol 35* (pp. 27730–27744). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

  ↩
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551. ↩
- Roberts, H., Bhargava, R., Valiukas, L., Jen, D., Malik, M. M., Bishop, C. S., Ndulue, E. B., Dave, A., Clark, J., Etling, B., Faris, R., Shah, A., Rubinovitz, J., Hope, A., D'Ignazio, C., Bermejo, F., Benkler, Y., & Zuckerman, E. (2021). Media cloud: Massive open source collection of blobal news on the open web. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 1034–1045. https://doi.org/10.1609/icwsm.v15i1.18127

  ↩
- Roberts, S. (2020, November 5). After the 2020 election, polling is dead. *Jacobin*. https://jacobin.com/2020/11/2020-election-polling-dead-biden-trump ↩
- Sanders, N. E., & Schneier, B. (2021). *Machine learning featurizations for AI hacking of political systems*. ArXiv. https://doi.org/10.48550/arXiv.2110.09231

  ↩
- Sanders, N. E., & Schneier, B. (2023a, January 15). How ChatGPT hijacks democracy. *The New York Times*. https://www.nytimes.com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy.html

  ↩
- Sanders, N. E., & Schneier, B. (2023b, March 14). How AI could write our laws. *Technology Review*. https://www.technologyreview.com/2023/03/14/1069717/how-ai-could-write-our-laws/

  ↩
- Sanders, N. E., & Schneier, B. (2023c, April 28). Just wait until Trump is a chatbot. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/04/ai-generated-political-ads-election-candidate-voter-interaction-transparency/673893/

  ↩

- Schaffner, B., Ansolabehere, S., & Shih, M. (2023). *Cooperative election study common content, 2022*. (Version V3) [Data set]. CCES Dataverse. https://doi.org/10.7910/DVN/PR4L8P

  ↩
- Schneider, D., & Harknett, K. (2022). What's to like? Facebook as a tool for survey data collection. *Sociological Methods & Research*, *51*(1), 108–140. https://doi.org/10.1177/0049124119882477 ↩
- Schneier, B. (2021, April). *The coming AI hackers*. Harvard Kennedy School, Belfer Center for Science and International Affairs. https://www.belfercenter.org/publication/coming-ai-hackers

  ↩
- Schneier, B., Farrell, Henry, & Sanders, N. E. (2023, April 21). How artificial intelligence can aid democracy. *Slate*. https://slate.com/technology/2023/04/ai-public-option.html ↩
- Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (NIST Special Publication 1270). National Institute of Standards and Technology, U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.1270

  ↩
- Silver, N. (2021, March 25). The death of polling is greatly exaggerated. *FiveThirtyEight*. https://fivethirtyeight.com/features/the-death-of-polling-is-greatly-exaggerated/ ↩
- Splichal, S. (2022). In data we (don't) trust: The public adrift in data-driven public opinion models. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517221097319

  ↩
- Stone, M., Aravopoulou, E., Ekinci, Y., Evans, G., Hobbs, M., Labib, A., Laughlin, P., Machtynger, J., & Machtynger, L. (2020). Artificial intelligence (AI) in strategic marketing decision-making: a research agenda. *The Bottom Line*, *33*(2), 183–200. https://doi.org/10.1108/BL-03-2020-0022 ↩
- Tonkiss, F. (2004). Using focus groups. In C. Seale (Ed.), *Researching society and culture*. Sage Publications.

  ↩
- Usher, K., Durkin, J., Martin, S., Vanderslott, S., Vindrola-Padros, C., Usher, L., & Jackson, D. (2021). Public sentiment and discourse on domestic violence during the COVID-19 pandemic in Australia: Analysis of social media posts. *Journal of Medical Internet Research*, *23*(10), Article e29025. https://doi.org/10.2196/29025

  ↩
- van der Wal, O., Bachmann, D., Leidinger, A., van Maanen, L., Zuidema, W., & Schulz, K. (2022). *Undesirable biases in NLP: Averting a crisis of measurement*. ArXiv. https://doi.org/10.48550/arXiv.2211.13709

  ↩

- Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In T. Linzen, G. Chrupała, Y. Belinkov, D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 63–76). https://doi.org/10.18653/v1/W19-4808

  ↩
- Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). *Voyager: An open-ended embodied agent with large language models*. ArXiv. https://doi.org/10.48550/arXiv.2305.16291 ↩
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems*. https://openreview.net/forum?id=_VjQlMeSB_J

  ↩
- Wu, P. Y., Tucker, J. A., Nagler, J., & Messing, S. (2023). *Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting*. ArXiv. https://doi.org/10.48550/arXiv.2303.12057 ↩
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). *Fine-tuning language models from human preferences*. ArXiv. https://doi.org/10.48550/arXiv.1909.08593 ↩