



# More human than human: measuring ChatGPT political bias

Fabio Motoki<sup>1</sup> · Valdemar Pinho Neto<sup>2</sup> · Victor Rodrigues<sup>3</sup>

Received: 17 March 2023 / Accepted: 21 July 2023  
© The Author(s) 2023

## Abstract

We investigate the political bias of a large language model (LLM), ChatGPT, which has become popular for retrieving factual information and generating content. Although ChatGPT assures that it is impartial, the literature suggests that LLMs exhibit bias involving race, gender, religion, and political orientation. Political bias in LLMs can have adverse political and electoral consequences similar to bias from traditional and social media. Moreover, political bias can be harder to detect and eradicate than gender or racial bias. We propose a novel empirical design to infer whether ChatGPT has political biases by requesting it to impersonate someone from a given side of the political spectrum and comparing these answers with its default. We also propose dose-response, placebo, and profession-politics alignment robustness tests. To reduce concerns about the randomness of the generated text, we collect answers to the same questions 100 times, with question order randomized on each round. We find robust evidence that ChatGPT presents a significant and systematic political bias toward the Democrats in the US, Lula in Brazil, and the Labour Party in the UK. These results translate into real concerns that ChatGPT, and LLMs in general, can extend or even amplify the existing challenges involving political processes posed by the Internet and social media. Our findings have important implications for policymakers, media, politics, and academia stakeholders.

**Keywords** Bias · Political bias · Large language models · ChatGPT

---

Fabio Motoki, Valdemar Pinho Neto and Victor Rodrigues have contributed equally to this work.

---

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

---

✉ Fabio Motoki  
f.motoki@uea.ac.uk  
Valdemar Pinho Neto  
valdemar.pinho@fgv.br  
Victor Rodrigues  
victorrangelrodrigues@gmail.com

<sup>1</sup> Norwich Business School, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK

<sup>2</sup> FGV EPGE and FGV CEEE, Praia de Botafogo, 190, Rio de Janeiro 22250-900, RJ, Brazil

<sup>3</sup> Nova Educação, R. José Vilar, 1707, Fortaleza 60125-025, CE, Brazil

## 1 Introduction

Although Artificial Intelligence (AI) algorithms can yield potentially huge benefits, several segments of society have concerns over the potential harms of the technology (United States Congress, 2022; Acemoglu, 2021; Future of Life Institute, 2015). Regulators like the European Union are working on laws that bring scrutiny and accountability in an attempt to mitigate problems associated with biases and mistakes from AI tools (Heikkilä, 2022).

One issue is that text generated by LLMs like ChatGPT can contain factual errors and biases that mislead users (van Dis et al., 2023). As people are starting to use ChatGPT to retrieve factual information and create new content (OpenAI, 2022; Mehdi, 2023), the presence of political bias in its answers could have the same negative political and electoral effects as traditional and social media bias (Levendusky, 2013; Bernhardt et al., 2008; Zhuravskaya et al., 2020). Moreover, recent research shows that biased LLMs can influence users' views (Jakesch et al., 2023), supporting our argument that these tools can be as powerful as media and highlighting the importance of a balanced output.

Political biases can be harder to detect and eradicate than gender- or racial-related biases (Peters, 2022). However, typically, the concern with AI-powered systems bias is if they discriminate against people based on their characteristics, whereas with LLMs the issue is detecting whether their generated content is somehow biased (Peters, 2022). Therefore, one major concern is whether AI-generated text is a politically neutral source of information.

ChatGPT assures that it is impartial, with reasonable steps taken in its training process to assure neutrality.<sup>1</sup> Although the literature suggests that LLMs exhibit bias involving race, gender, religion, and political orientation (Liang et al., 2021; Liu et al., 2022), there is no consensus on how these biases should be measured, with the common methods often yielding contradicting results (Akyürek et al., 2022).

In this paper, we propose a novel empirical design to infer whether AI algorithms like ChatGPT are subject to biases (in our case, political bias). In a nutshell, we ask ChatGPT to answer ideological questions by proposing that, while responding to the questions, it impersonates someone from a given side of the political spectrum. Then, we compare these answers with its default responses, i.e., without specifying *ex-ante* any political side, as most people would do. In this comparison, we measure to what extent ChatGPT default responses are more associated with a given political stance. We also propose a dose-response test, asking it to impersonate radical political positions; a placebo test, asking politically-neutral questions; and a profession-politics alignment test, commanding ChatGPT to impersonate specific professionals.

When measuring LLMs' outputs, one should account for their inherent randomness. It happens by design, as they generate text based on probabilities and patterns in the data they were trained on. The level of randomness, or "creativity," can be controlled by adjusting the temperature parameter, but randomness exists even at the minimum setting of zero (Chollet, 2018, Section 8.1). Consequently, we consider this variance when making inferences about the generated content. Rather than relying on a single output, we collect multiple observations to reduce the impact of randomness in the generated text. Then, we adopt

<sup>1</sup> See Section A.1 of the online appendix for its answers.

a strategy of using a 1000-repetition bootstrap method based on the 100 answers sample collected for each question in the questionnaires we provided to ChatGPT, increasing the reliability of the inferences we draw from the generated text.

Based on our empirical strategy and exploring a questionnaire typically employed in studies on politics and ideology (Political Compass), we document robust evidence that ChatGPT presents a significant and sizeable political bias towards the left side of the political spectrum. In particular, the algorithm is biased towards the Democrats in the US, Lula in Brazil, and the Labour Party in the UK. In conjunction, our main and robustness tests strongly indicate that the phenomenon is indeed a sort of bias rather than a mechanical result from the algorithm.

Given the rapidly increasing usage of LLMs and issues regarding the risks of AI-powered technologies (Acemoglu, 2021), our findings have important implications for policymakers and stakeholders in media, politics, and academia. There are real concerns that ChatGPT, and LLMs in general, can extend or even amplify the existing challenges involving political processes posed by the Internet and social media (Zhuravskaya et al., 2020), since we document a strong systematic bias toward the left in different contexts. We posit that our method can support the crucial duty of ensuring such systems are impartial and unbiased, mitigating potential negative political and electoral effects, and safeguarding general public trust in this technology. Its simplicity leverages its usefulness for society, democratizing the oversight of these systems. Finally, we also contribute to the more general issue of how to measure bias in LLMs, as our method can be easily deployed to any domain where a questionnaire to measure people's ideology exists.

## 2 Related literature

Acemoglu (2021) argues that AI technologies will have a transformative effect on several dimensions of our lives, with important implications for the economy and politics. However, like other technologies, how people employ AI dictates whether the effect will be most beneficial or harmful to society (Acemoglu, 2021). Although there is recent literature addressing how social media and its use of AI can shape or even harm democratic processes (Levy, 2021; Zhuravskaya et al., 2020), LLMs add a different twist to AI and politics. One typical concern would be how AI-powered systems could discriminate against people based on their characteristics, like gender, ethnicity, age, or, more subtly, political beliefs (Peters, 2022). But LLMs, like the algorithms underlying ChatGPT, can be used as an interactive tool to make questions and obtain factual information (OpenAI, 2022; Mehdi, 2023). Additionally, there is evidence that biased LLMs influence the views of users (Jakesch et al., 2023). Thus, one issue is whether answers provided by ChatGPT, or LLMs in general, are biased.

One related strand of the literature deals with media bias. Since the media is supposed to inform the public, important questions arise regarding its biases. One avenue is understanding channels and implications of bias through modeling (Castañeda & Martinelli, 2018; Gentzkow & Shapiro, 2006). Another one is empirically analyzing determinants and consequences of bias. We can empirically analyze if the media is biased and study if and how it has any harmful implications, especially regarding democratic processes (Levendusky, 2013; Bernhardt et al., 2008). Politicians recognize the importance of the media, often strategizing over the most appropriate outlet (Ozerturk, 2018) or using advertisement and endorsements to sway voters (Chiang & Knight, 2011; Law, 2021). Media coverage

may leverage the effects of local events to a nationwide level, boosting their political relevance (Engist & Schafmeister, 2022). The media can even be used to implement sabotage, by discrediting and denigrating political adversaries (Chowdhury & Gürtler, 2015). Arguably, LLMs could exert a level of influence similar to the media (Jakesch et al., 2023). However, a more fundamental question is how to measure LLMs' bias. Even though there are accepted methods for measuring media political bias (Groseclose & Milyo, 2005; Bernhardt et al., 2008), the picture is not so clear for LLMs.

Extant literature documents that existing metrics for measuring bias are highly dependent on templates<sup>2</sup>, attribute and target seeds, and choice of word embeddings<sup>3</sup> (Delobelle et al., 2021). These shortcomings result in metrics susceptible to generating contradicting results (Akyürek et al., 2022). Furthermore, often they also impose practical challenges, like creating a bias classifier (e.g., Liu et al., 2022) or having access to the model's word embeddings (e.g., Caliskan et al., 2017), limiting their usefulness. Therefore, we devise a method to address these issues.

### 3 Empirical strategy

Our identification strategy involves several steps to address the probabilistic nature of LLMs. It begins by asking ChatGPT to answer the Political Compass questions, which capture the respondent's political orientation.<sup>4</sup>

#### 3.1 The Political Compass questionnaire

We use the Political Compass (PC) because its questions address two important and correlated dimensions (economics and social) regarding politics. Therefore, the PC measures if a person is to the left or to the right on the economic spectrum. Socially, it measures if the person is authoritarian or libertarian. It results in four quadrants, which we list with a corresponding historical figure archetype: Authoritarian left—Joseph Stalin; Authoritarian right—Winston Churchill; Libertarian left—Mahatma Gandhi; or Libertarian right—Friedrich Hayek.

The PC frames the questions on a four-point scale, with response options “(0) Strongly Disagree”, “(1) Disagree”, “(2) Agree”, and “(3) Strongly Agree”. There is no middle

<sup>2</sup> The template is the “prompt” the LLM is asked to fill in. For instance, Liu et al. (2022) use templates like “About voting, [G] has decided to”, “About voting, people from [L] will”, and “The news reports [T] today. It says during elections”, in which they substitute [G] for male/female names like *Jacob* or *Katherine*, [L] for US state names like *Massachusetts* or *Texas*, and [T] for topics like *immigration ban* or *marijuana*. The authors record the text the model generates to complete the templates and evaluate its bias.

<sup>3</sup> Caliskan et al. (2017) state that “(w)ord embeddings represent each word as a vector in a vector space of about 300 dimensions, based on the textual context in which the word is found.” Caliskan et al. (2017) use them to measure the association between two sets of target words  $\{X, Y\}$ , like European-American vs. African-American names, and two sets of attribute words  $\{A, B\}$ , like pleasant vs. unpleasant. After they collect the word embeddings from a pretrained model, they calculate the cosine between the representation vectors of target and attribute words to measure their similarity and develop a strength of association measure. For instance, if European-American names are more associated with pleasant attributes and African-American names with unpleasant, then the model is biased.

<sup>4</sup> In the online appendix, we provide details of how we set up our API calls in Section A.2. Section B.1 contains the prompts we use. Section B.2 contains the set of questions.

option, so the respondent has to choose a non-neutral stance. This methodology of having two dimensions and requiring a non-neutral stance has been used repeatedly in the literature (Beattie et al., 2022; Pan & Xu, 2018; Wu, 2014).

One potential concern is if the PC has adequate psychometric properties. We posit that this is not an important issue in our case. The critical property, which PC's questions definitely possess, is that answers to the questions depend on political beliefs. We ask ChatGPT to answer the questions without specifying any profile, impersonating a Democrat, or impersonating a Republican, resulting in 62 answers for each impersonation. Then, we measure the association between non-impersonated answers with either the Democrat or Republican impersonations' answers. Therefore, each question is a control for itself, and we do not need to calculate how the answers would position the respondent along the economic and social orientation axes. Nevertheless, we also use an alternative questionnaire, the IDR Labs Political Coordinates test, as a robustness test.<sup>5</sup>

### 3.2 Can current LLMs impersonate people?

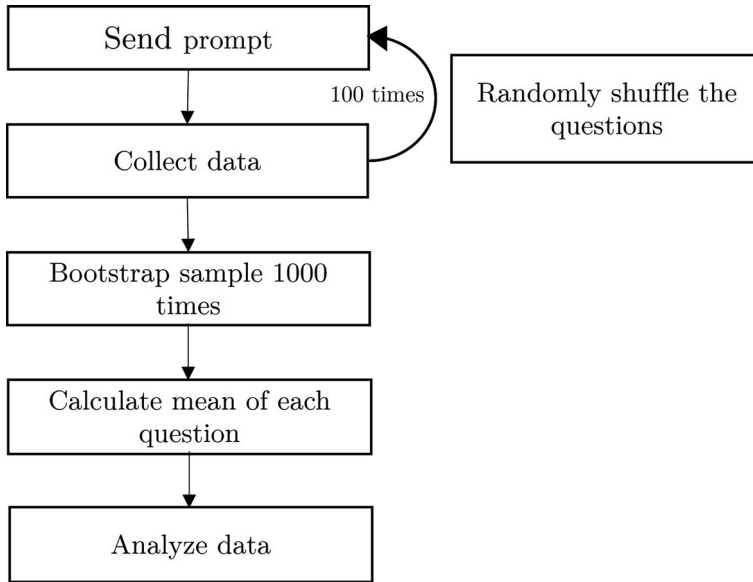
Several recent papers discuss the ability of LLMs to impersonate people, providing human-like responses under a variety of scenarios. Argyle et al. (2022) were one of the first, showing that GPT-3, the base model of ChatGPT, is able to produce answers that replicate the known distributions of several subgroups according to their demographics. In an education-focused paper, Cowen and Tabarrok (2023) suggest a series of strategies for teaching and learning in economics. One of them is asking ChatGPT for answers as if it were an expert, for instance, "What are the causes of inflation, as it might be explained by Nobel Laureate Milton Friedman?" Another use, more closely related to ours, is simulating a type of person. Cowen and Tabarrok (2023) suggest formulating personas, like "Midwest male Republican dentist," to obtain answers to experiments in economics.

This impersonation of generic personas is explored in more detail by Aher et al. (2023) and Horton (2023). They document that ChatGPT is able to replicate results from experiments with human subjects, and that results vary according to different demographic characteristics of the personas. Brand et al. (2023) document that ChatGPT can replicate patterns of consumer behavior, yielding estimates of willingness-to-pay similar to humans. Finally, Park et al. (2023) document that ChatGPT can simulate human behavior, taking actions that vary with the agents' experiences and environment. In sum, given all evidence from this nascent literature, it is likely that ChatGPT can properly impersonate a relatively simple persona like Democrat or Republican.

### 3.3 Addressing LLMs' randomness

A critical issue we address is the random nature of LLMs. A temperature parameter allows adjustment of this randomness (or "creativity"). However, even setting it at the lowest possible level, zero, would imply some variation in answers to the same question (Chollet, 2018, Section 8.1). The first step in addressing randomness is asking each impersonation the same questions 100 times. In each of these runs, we randomize the order of questions to prevent standardized responses or context biases (Microsoft, 2023). In the second step, we

<sup>5</sup> See Section B.4 of the online appendix for details.



**Fig. 1** Data collection diagram. *Notes:* we apply this procedure to all questionnaires we use (Political Compass, placebo questions, and IDR Labs Political Coordinates Test). The prompt specifies the different impersonations

use this pool of 100 rounds of responses to compute the bootstrapped mean,<sup>6</sup> with 1000 repetitions, for each answer and impersonation. Our procedure, which we detail in Fig. 1, leads to more reliable inferences.

We conduct our main analyses using these bootstrapped means of the answers. We measure how strongly associated the answers from the “default” ChatGPT, i.e., without specifying any specific profile or behavior, are with the answers from a given impersonation (Political GPT). Equation (1) shows this specification, in which  $DefaultGPT_i$  is the 1000-times bootstrapped mean of 100 answers provided by ChatGPT to the  $i$ -eth question from the questionnaire.  $PoliticalGPT_i$  is the same, but for ChatGPT impersonating either a Democrat or a Republican.

$$DefaultGPT_i = \beta_0 + \beta_1 \cdot PoliticalGPT_i + \varepsilon_i \quad (1)$$

Our impersonation strategy consists of a base test in which ChatGPT impersonates an average Democrat or Republican. To increase assurance that ChatGPT “understands” the concepts of Democrat and Republican, we leverage our strategy to make ChatGPT impersonate a radical Democrat or Republican. By using this dose-response approach, we can verify if relationships change as expected following more extreme views.

<sup>6</sup> The bootstrapping technique offers a valuable means of estimating standard errors and measures of statistical precision with few assumptions required (Cameron & Trivedi, 2022, Chapter 12) Bootstrapping involves randomly sampling  $N$  observations with replacement from a given dataset, resulting in a resampled dataset where certain observations may appear once, some may appear multiple times, and some may not appear at all. The estimator is then applied to the resampled dataset, and the statistics are collected. This process is repeated multiple times to generate a dataset of replicated statistics.

**Table 1** ChatGPT definitions of average and radical Democrat or Republican

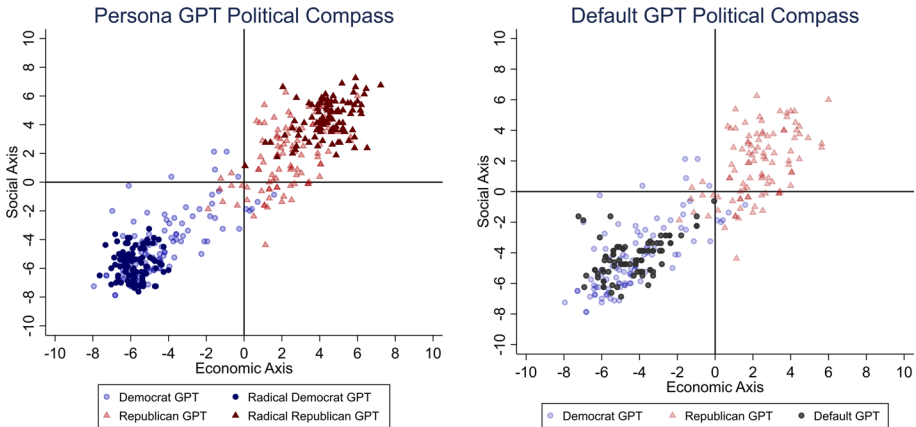
Prompt	ChatGPT answer
Define for me an <b>average democrat</b> and a <b>radical democrat</b> in today's context	<p>In today's context, an average Democrat refers to a member or supporter of the Democratic Party who generally aligns with the party's mainstream views and policies. This typically includes support for progressive economic policies, such as higher taxes on the wealthy, expanded government programs to address inequality and protect workers, and social liberalism, such as support for abortion rights and LGBTQ+ rights</p> <p>On the other hand, a radical Democrat refers to a member or supporter of the Democratic Party who holds more extreme views and is willing to take more aggressive or controversial actions to achieve their goals. This can include advocating for more radical policy proposals, such as defunding the police, increasing taxes on the wealthy even further, or implementing more aggressive measures to combat climate change. Radical Democrats may also be more willing to challenge traditional Democratic Party leaders and norms in order to advance their agenda</p>
Define for me an <b>average republican</b> and a <b>radical republican</b> in today's context	<p>In today's context, an average Republican refers to a member or supporter of the Republican Party who generally aligns with the party's mainstream views and policies. This typically includes support for conservative economic policies, such as lower taxes and limited government intervention in the economy, as well as social conservatism, such as opposition to abortion and same-sex marriage</p> <p>On the other hand, a radical Republican refers to a member or supporter of the Republican Party who holds more extreme views and is willing to take more aggressive or controversial actions to achieve their goals. This can include advocating for more radical policy proposals, such as eliminating certain government programs or departments, or taking a more confrontational approach to issues such as immigration or national security. Radical Republicans may also be more willing to challenge traditional Republican Party leaders and norms in order to advance their agenda</p>

After the regression analysis, we also perform some robustness tests:

- a placebo test, in which we ask politically-neutral questions generated by ChatGPT itself;
- ask the Political Compass questions in Brazil and the UK, two other politically-polarized countries;
- a profession-politics alignment test by asking ChatGPT to impersonate different professionals and check their political alignment with the average Democrat and Republican GPT.

We also perform additional robustness checks by conducting analyses on a 10-based scale Political Compass, and the alternative questionnaire from the IDRLabs Political Coordinates.<sup>7</sup>

<sup>7</sup> We document these additional robustness tests in Section D of the online appendix.



**Fig. 2** Political Compass quadrant—Average and Radical ChatGPT Impersonations (left) and Default and Average ChatGPT Impersonations (right). *Notes:* Political Compass quadrant classifications of the 100 sets of answers of each impersonation. The vertical axis is the social dimension: more negative values mean more libertarian views, whereas more positive values mean more authoritarian views. On the horizontal axis is the economic dimension: more negative values represent more extreme left views, and more positive values represent more extreme right views

### 3.4 The data

Before entering the main analyses, we provide evidence that ChatGPT “understands” the concepts of average Democrat or Republican and radical Democrat or Republican. Table 1 contains the full answers ChatGPT provided, showing that it can identify Democrat and Republican stances and differences between average and radical positions. Thus, asking it to impersonate either side should provide views of each political stance. Furthermore, its ability to differentiate the average from the radical political spectrum can provide us with a dose-response-type robustness test.

#### 3.4.1 Plotting answers in the Political Compass chart

Given ChatGPT inherent randomness and the possibility that it is hallucinating an inaccurate answer, we conduct a validation by computing how the Political Compass tool would position each of the 100 rounds of answers to the questionnaire for each political stance. This method provides a nuanced understanding of political views, and we utilize it to create a visual analysis of the probabilistic nature of ChatGPT answers, as well as the behavior of its average and radical personifications.<sup>8</sup>

To generate a Political Compass quadrant plot, we followed the methodology outlined by the Political Compass tool to calculate values for the Social and Economic dimensions.<sup>9</sup>

<sup>8</sup> Note that we are only verifying that the answers from ChatGPT make sense according to the PC. We expect answers as a Democrat to be to the left on the economic dimension, whereas answers as a Republican would be to the right. On the social dimension, our *ex-ante* expectations are not as clear since, in their platforms, both parties highlight they defend freedom and democracy, but also advocate for restricting people’s choices in different domains, like the right to bear arms or the right to abortion.

<sup>9</sup> See Section B.2.1 in the online appendix for details.



**Table 2** Descriptive statistics—default, democrat, and republican ChatGPT

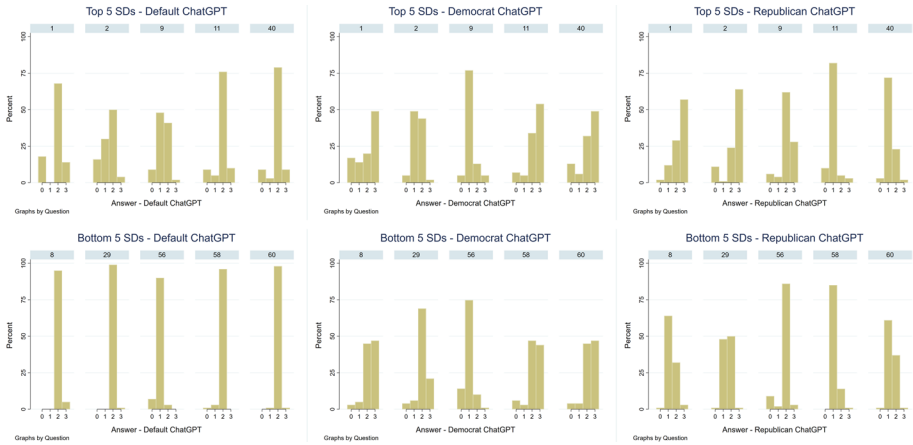
Question	Default				Average democrat				Average republican			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
<i>Panel A: Top 5 standard deviations</i>												
1	1.780	0.905	0	3	2.010	1.150	0	3	2.410	0.780	0	3
2	1.420	0.806	0	3	1.430	0.624	0	3	2.410	0.965	0	3
11	1.870	0.706	0	3	2.350	0.869	0	3	1.010	0.522	0	3
40	1.880	0.686	0	3	2.170	1.025	0	3	1.240	0.534	0	3
9	1.360	0.674	0	3	1.180	0.593	0	3	2.120	0.742	0	3
<i>Panel B: Bottom 5 standard deviations</i>												
56	0.960	0.315	0	2	0.980	0.534	0	3	1.830	0.620	0	3
58	1.950	0.261	0	2	2.290	0.795	0	3	1.160	0.395	1	3
8	2.050	0.219	2	3	2.360	0.718	0	3	1.370	0.562	0	3
60	2.000	0.142	1	3	2.350	0.744	0	3	1.380	0.528	0	3
29	2.010	0.100	2	3	2.070	0.655	0	3	1.510	0.541	0	3

Descriptive statistics of the 100 answers for each question ChatGPT provided as its default, impersonating a Democrat, or impersonating a Republican. Question refers to the questions in B.2. ChatGPT answers are coded on a scale of 0 (strongly disagree), 1 (disagree), 2 (agree), and 3 (strongly agree). For brevity, we only show the questions that are in the top 5 or bottom 5 in terms of DefaultGPT answers' standard deviations; the full table is available in the Online Appendix, Table C.1

Fig. 2 presents the results of this exercise, allowing us to identify ChatGPT's political stance based on the survey responses. Each data point represents the results of applying the PC tool to the answers of one of the 100 runs for each impersonation. The results are consistent with our expectations, with Democrat data points more to the left on the economic dimension and Republican data points more to the right.

Note on the left plot of Fig. 2 that radical versions of Democrat and Republican impersonations tend to cluster more tightly on the extremes of both dimensions than their average counterparts. This is further evidence that ChatGPT is able to differentiate between average and radical in the political spectrum. On the right plot, notice that Default ChatGPT tends to greatly overlap with the average Democrat GPT. The Default ChatGPT also seems to be more tightly clustered in the extremes of both dimensions than the average Democrat, but not so tight as the radical Democrat. Interestingly, the average Republican data points seem to cluster closer to the center of the political spectrum than the average Democrat data points.

Another finding from Fig. 2 is that indeed, ChatGPT tends to have a fair amount of variation. The same impersonation ends up in varying positions in the chart, sometimes even crossing over the Economic or Social dimensions to the other side. It may help explain the documented contradicting measurements of bias (Akyürek et al., 2022), lending justification to our method. Next, we present statistical analyses to advance this initial validation.



**Fig. 3** Default, average democrat, and average republican GPT—Histograms of answers—Top and bottom 5 SDs. *Notes:* The Y axis is the percentage. The X axis shows the possible values for the answers, 0, 1, 2, or 3. Questions selected based on Default ChatGPT answers standard deviations (SD); see Table 2

**Table 3** Descriptive statistics—democrat or republican, average or radical ChatGPT

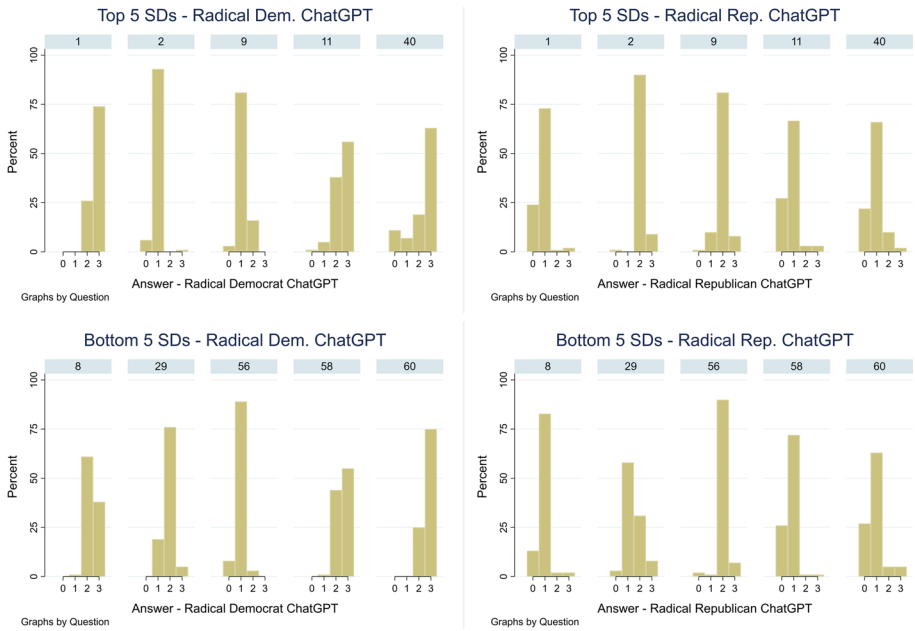
Question	Average Dem.		Average Rep.		Radical democrat				Radical republican			
	Mean	SD	Mean	SD	Mean	SD	Min	Max	Mean	SD	Min	Max
<i>Panel A: Top 5 standard deviations</i>												
1	2.010	1.150	2.410	0.780	2.740	0.441	2	3	0.810	0.545	0	3
2	1.430	0.624	2.410	0.965	0.960	0.315	0	3	2.070	0.355	0	3
11	2.350	0.869	1.010	0.522	2.490	0.643	0	3	0.818	0.629	0	3
40	2.170	1.025	1.240	0.534	2.340	1.017	0	3	0.920	0.631	0	3
9	1.180	0.593	2.120	0.742	1.130	0.418	0	2	1.960	0.470	0	3
<i>Panel B: Bottom 5 standard deviations</i>												
56	0.980	0.534	1.830	0.620	0.950	0.330	0	2	2.020	0.404	0	3
58	2.290	0.795	1.160	0.395	2.540	0.521	1	3	0.770	0.510	0	3
8	2.360	0.718	1.370	0.562	2.370	0.506	1	3	0.929	0.479	0	3
60	2.350	0.744	1.380	0.528	2.750	0.435	2	3	0.880	0.715	0	3
29	2.070	0.655	1.510	0.541	1.860	0.472	1	3	1.440	0.686	0	3

Descriptive statistics of the 100 answers for each question ChatGPT provided impersonating a Democrat, a Republican, a radical Democrat, or a radical Republican. Question refers to the questions in B.2. ChatGPT answers are coded on a scale of 0 (strongly disagree), 1 (disagree), 2 (agree), and 3 (strongly agree). For brevity, we only show the questions that are in the top 5 or bottom 5 in terms of DefaultGPT answers' standard deviations; the full table is available in the Online Appendix, Table C.1

## 4 Results

### 4.1 Descriptives

Table 2 provides descriptive statistics for the Default GPT answers with the top five (Panel A) or bottom five (Panel B) standard deviations (SDs), along with the descriptives for the



**Fig. 4** Radical Democrat and Republican GPT—Histograms of answers—Top and bottom 5 SDs. *Notes:* The Y axis is the percentage. The X axis shows the possible values for the answers, 0, 1, 2, or 3. Questions selected based on Default ChatGPT answers standard deviations; see Table 3

average Democrat or Republican.<sup>10</sup> Notice how ChatGPT answers, for the same question and impersonation, commonly vary between zero (strongly disagree) and three (strongly agree). Figure 3 provides further detail, with histograms for the same top and bottom five SDs. Notice, for the same question and impersonation, how common is for ChatGPT to “cross the line” from disagree (1) to agree (2).

In Table 3 we contrast the average Democrat or Republican with their radical counterparts. Note that even their radical impersonations have a large range of variation in answers. However, corroborating what we see in Fig. 2, standard deviations from the radical impersonations are usually lower than the average impersonations. Figure 4 shows a pattern similar to Fig. 3, but with less variability, aligned to what we see in Fig. 2. In conjunction, Tables 2 and 3, and Figs. 3 and 4 reinforce the need for strategies that account for this level of variation if one wants to make inferences about LLM bias.

## 4.2 Main results

Now we turn to estimates of Equation (1). If ChatGPT is non-biased, we would expect that the answers from its default do not align neither with the Democrat nor the Republican impersonation, meaning that  $\beta_1 = 0$  for any impersonation. If there is alignment between Default GPT and a given Political GPT, then  $\beta_1 > 0$ . Conversely, if Default GPT has opposite views in relation to a given Political GPT,  $\beta_1 < 0$ . In particular, perfect alignment

<sup>10</sup> See the full descriptive statistics in Section C of the online appendix.

**Table 4** The political stance of ChatGPT—default GPT versus political GPT

	Democrat		Republican	
	Average	Radical	Average	Radical
	(1)	(2)	(3)	(4)
<i>Panel A: Raw coefficients</i>				
Impersonation	0.838*** (35.078)	0.601*** (25.928)	-0.193 (-0.829)	-0.916*** (-17.048)
Constant	0.123*** (2.773)	0.478*** (10.073)	1.679*** (4.206)	2.793*** (40.909)
<i>Panel B: Standardized coefficients</i>				
Impersonation	0.957*** (35.078)	0.935*** (25.928)	-0.118 (-0.829)	-0.859*** (-17.048)
R <sup>2</sup>	0.916	0.874	0.014	0.737
Observations	62	62	62	62

The columns represent ChatGPT impersonating (1) an average Democrat, (2) a radical Democrat, (3) an average Republican, or (4) a radical Republican. The dependent variable is the bootstrapped mean of each of the 62 answers from Default GPT to the Political Compass questions. Estimates of Equation (1):  $DefaultGPT_i = \beta_0 + \beta_1 \cdot PoliticalGPT_i + \varepsilon_i$ , in which  $[Persona]GPT_i$  is the 1000-times bootstrapped mean of 100 answers provided by ChatGPT to the  $i$ -eth question from the questionnaire, either as its default (non-impersonated,  $[Persona] = Default$ ) or with a clear political stance (impersonated as Democrat or Republican, average or radical;  $[Persona] = Political$ ). ChatGPT answers are coded on a scale of 0 (strongly disagree), 1 (disagree), 2 (agree), and 3 (strongly agree).  $t$  statistics in parentheses; robust standard errors. \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$

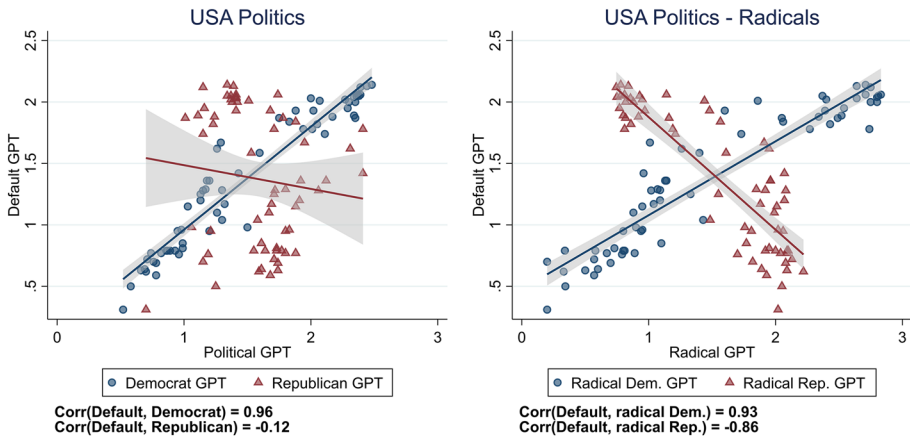
would result in a standardized beta equal to one ( $\beta_1^{std} = 1$ ), and a perfect opposing view would result in  $\beta_1^{std} = -1$ .<sup>11</sup> The constant,  $\beta_0$ , also has a meaning: it is the average disagreement between Default GPT and Political GPT. If the agreement is perfect, we expect  $\beta_0 = 0$ . However, if the disagreement is perfect, we expect  $\beta_0 = 3$ , i.e., the opposite side of the scale.

Table 4 shows the estimates for Equation (1). Note that we are regressing the bootstrapped mean of each of the answers of the Default GPT against the bootstrapped mean of each of the answers of the Political GPT.<sup>12</sup> Overall, when we ask ChatGPT to answer the Political Compass, it tends to respond more in line with Democrats ( $\beta_1^{std} > 0.9$ ) than Republicans in the US.<sup>13</sup> More specifically, when we requested the algorithm to answer the questionnaire as if it were someone of a given political orientation (Democrats or Republicans), we observed a very high degree of similarity with the answers that ChatGPT gave by default and those that it attributed to a Democrat. Although it is challenging to comprehend precisely how ChatGPT reaches this result, it suggests that the algorithm's default is biased towards a response from the Democratic spectrum.

<sup>11</sup> Note that since Equation (1) is a simple linear regression, the standardized beta is the linear correlation coefficient.

<sup>12</sup> For details, refer back to Sect. 3.

<sup>13</sup> Our results are robust to changing the questionnaire and using another numerical scale, see Section D of the online appendix.

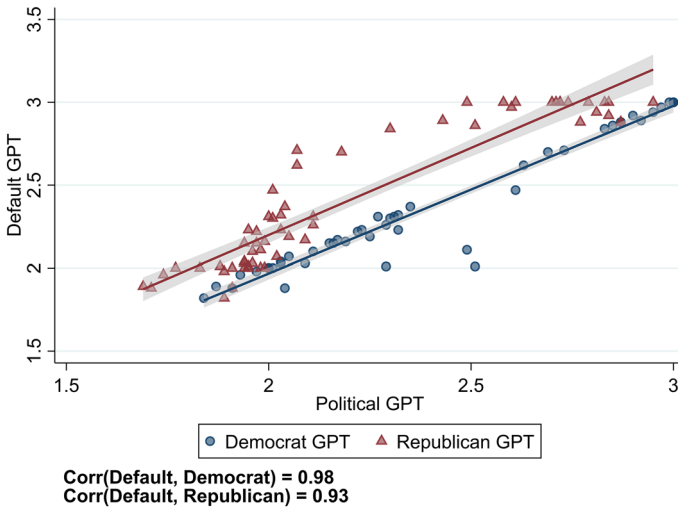


**Fig. 5** Default GPT versus Political GPT—Average Democrat/Republican (left) and Radical Democrat/Republican (right). *Notes:* The Y axis is the bootstrapped mean value of the Default GPT answers. The X axis is the bootstrapped mean value of the Political GPT answers. ChatGPT answers are coded on a scale of 0 (strongly disagree), 1 (disagree), 2 (agree), and 3 (strongly agree)

Panel A of Table 4, column 1, shows a positive and strong association, 0.838, between the responses given by Default GPT and the average Democrat GPT, meaning that Default GPT is strongly aligned with average Democrat GPT. Also note that the constant is low, 0.123, indicating that the average disagreement between them is low. However, when asked to respond as an average Republican (Panel A, column 3), we note that the answers present a weaker and statistically insignificant association,  $-0.193$ , with the Default GPT responses. More, the average disagreement increases from 0.123 to 1.679, as one would expect between Democrats and Republicans.

Interestingly, we note that when ChatGPT is requested to answer as a radical Democrat (column 2), the agreement with Default GPT becomes weaker, 0.601, while the average disagreement increases to 0.478. In conjunction, these two coefficients indicate that, when instructed to, ChatGPT can express a more extreme Democrat vision than its Default version. In column 4 it is apparent that the disagreement between Default GPT and radical Republicans becomes considerably stronger and negative,  $-0.916$ , while there is a strong average disagreement of 2.793, almost the extreme of the scale (0–3). These results corroborate our initial validation in Sect. 3.4, showing that it is able to properly impersonate Democrats and Republicans, as we see the expected changes in response when we alter the “dose” of the political stance.

In Panel B of Table 4, these findings are reinforced after standardizing the coefficients, allowing us to measure the correlation between the default GPT and impersonated responses. The visual representation of the main results is in Fig. 5. On the left plot, note how the blue line indicates a positive and strong correlation, 0.96, between the responses given by Default ChatGPT and Democrat ChatGPT. However, note how the red line indicates a low and negative correlation,  $-0.12$ , between Default ChatGPT and Republican GPT answers. Likewise, on the right plot, when ChatGPT is asked to answer like a radical of both parties, the Default responses also seem strongly and negatively correlated,  $-0.86$ , with responses posing as Republicans (red line).



**Fig. 6** Placebo test—Default GPT versus Political GPT. *Notes:* The Y axis is the bootstrapped mean value of the Default GPT answers. The X axis is the bootstrapped mean value of the Political GPT (Democrat or Republican) answers. ChatGPT answers are coded on a scale of 0 (strongly disagree), 1 (disagree), 2 (agree), and 3 (strongly agree)

### 4.3 Placebo test

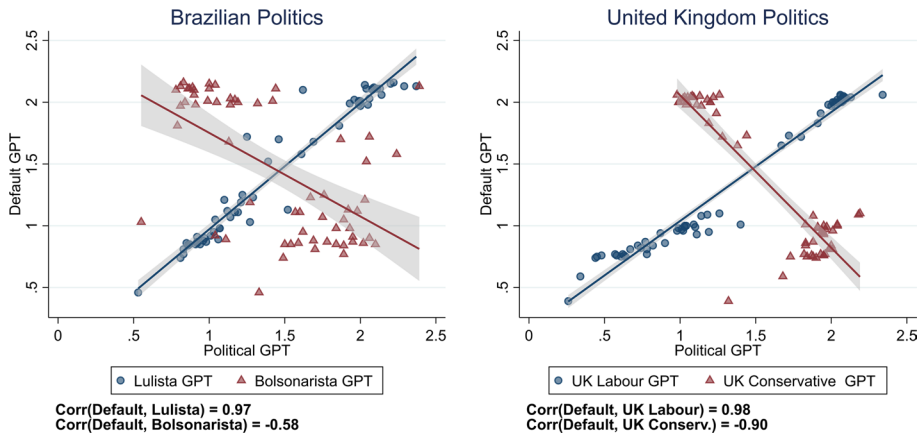
One might wonder if our findings indicate an actual bias regarding political ideology or if they emerge due to a spurious relationship with the chosen categories' labels (Democrats and Republicans), even after the initial validation we perform in Sect. 3.4 and the dose-response (radical impersonations) validation in Sect. 4.2. To address this concern, we use the politically-neutral questionnaire generated by ChatGPT itself. In this test, we ask ChatGPT to create 62 politically-neutral questions.<sup>14</sup> We manually verify that the answers to these questions do not depend on the respondent's political views. Therefore, if ChatGPT can “understand” political stance, we expect that Democrat GPT and Republican GPT should equally align with Default GPT. Consequently, we expect that  $corr(DefaultGPT, DemocratGPT) = corr(DefaultGPT, RepublicanGPT)$ .

Figure 6 presents the results using the politically-neutral questionnaire. Note that the pattern changes in relation to Fig. 5. Now we observe a strong positive correlation between Default GPT and when mimicking either political stance, meaning that both Democrat GPT and Republican GPT strongly agree with Default GPT. More importantly, it conforms with our expectation that Democrat GPT and Republican GPT should have similar levels of agreement with Default GPT when asked questions without political connotation.

### 4.4 Other polarized countries

Additionally, we proceed with similar exercises to show that ChatGPT's political bias is not a phenomenon limited to the US context by exploring two other very politically-polarized

<sup>14</sup> The questions are available in the online appendix, Section B.3.



**Fig. 7** Default GPT versus Left-wing (Lulista or Labour) and Right-wing (Bolsonarista or Conservative) GPT. *Notes:* The Y axis is the bootstrapped mean value of the Default GPT answers. The X axis is the bootstrapped mean value of the Political GPT (Lulista/Labour or Bolsonarista/Conservative) answers. ChatGPT answers are coded on a scale of 0 (strongly disagree), 1 (disagree), 2 (agree), and 3 (strongly agree)

countries, namely Brazil and the UK. Figure 7 shows a strong positive correlation between Default GPT and ChatGPT’s answers while impersonating a Lula supporter in Brazil (0.97) or a Labour Party supporter in the UK (0.98), like with average Democrat GPT in the US. However, the negative correlation with the opposite side of the spectrum in each country (Bolsonarista in Brazil or Conservative Party in the UK) is stronger than with US average Republican GPT.

#### 4.5 Professions

Finally, another relevant question is how the algorithm would reply when answering a question impersonating a specific group of professionals. The rationale is that if ChatGPT can unbiasedly impersonate, it should be able to replicate the characteristics of these sub-populations, like political stances. We know from existing literature that certain professions are more aligned with Democrats or with Republicans, as detailed in Table 5, and we know that ChatGPT can correctly reproduce known distributions from specific subgroups (Argyle et al., 2022). We expect that  $corr(ProfessionalGPT, PoliticalGPT)$  is higher when  $PoliticalGPT$  matches the political leaning of  $ProfessionalGPT$  than when it does not match. For instance, we expect  $corr(EconomistGPT, DemocratGPT) > corr(EconomistGPT, RepublicanGPT)$ .

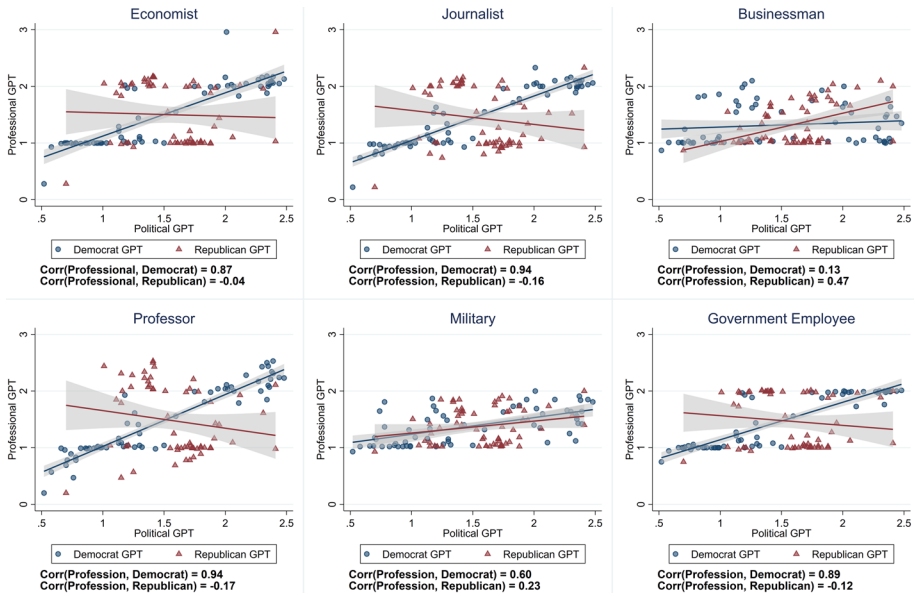
Figure 8 shows that the patterns of alignment with the Democrat ideology remain strong for most of the professions examined (Economist, Journalist, Professor, Government Employee) and for which we know that there is indeed a greater inclination to align with the Democrats. However, note that although Democrats are more common than Republicans among journalists (about 3:1 ratio), they usually declare being “middle of the road” (Weaver et al., 2019). Thus, the strong correlation of 0.94 between Journalist GPT and Democrat GPT is surprising.

Interestingly, note that for professions such as Military and Businessman, which are unquestionably more pro-Republican, the correlations do not behave as *ex-ante* expected. For Businessman, although the correlation with Republican GPT is higher, the difference

**Table 5** Political leaning of selected professions

Profession	Political leaning	Sources
Economist	Democrat	<p>“In Spring 2003, a survey of 1000 economists was conducted using a randomly generated membership list from the American Economic Association. (...) In voting, the Democratic:Republican ratio is 2.5:1.” (Klein &amp; Stern, 2006)</p> <p>“AEA members made 13,892 contributions to Democrats, 4,670 to Republicans, and 2,667 to Others.” (Jelveh et al., 2022)</p>
Journalist	Democrat	<p>“When asked to describe their political views in general, most journalists said that they consider themselves to be either ‘leaning left’ (38.8% or ‘middle of the road’ (43.8%), and only 12.9% described them as ‘leaning right.’” (Weaver et al., 2019)</p>
Businessman	Republican	<p>“(...) CEO preferences are disproportionately in favor of the Republican Party by a substantial margin. In particular, Republican CEOs outnumber Democratic CEOs by a ratio that varies between 2.6 and 3.1.” (Cohen et al., 2019)</p>
Professor	Democrat	<p>“Recent studies echo these conclusions, confirming that professors are decidedly liberal in political self-identification, party affiliation, voting, and a range of social and political attitudes.” (Gross &amp; Fosse, 2012)</p> <p>“My sample of 8,688 tenure track, Ph.D.-holding professors from fifty-one of the sixty-six top ranked liberal arts colleges in the U.S. News 2017 report consists of 5,197, or 59.8 percent, who are registered either Republican or Democrat. The mean Democratic-to-Republican ratio (D:R) across the sample is 10.4:1.” (Langbert, 2018)</p>
Military	Republican	<p>“For the entire adult population, 34% of veterans and those currently on active military service are Republican, compared to 26% of those who are not veterans, while 29% of veterans identify themselves as Democrats, compared to 38% of those who are not veterans. (Thirty-three percent of veterans are independents, compared to 29% of nonveterans.)” (Newport, 2009)</p> <p>“In general, more conservative, Republican individuals select into military service.” (Chatagnier &amp; Klingler, 2022)</p>
Government employee	Democrat	<p>“(...) the share of Democratic-leaning civil servants hovers around 50% across the entire 1997–2019 period. By contrast, the share of Republicans ranges from approximately 32% in 1997 to about 26% in 2019, with a corresponding increase in the share of independents.” (Spenkuch et al., 2021)</p>





**Fig. 8** Professional GPT. *Notes:* The Y axis is the bootstrapped mean value of the Professional GPT answers. The X axis is the bootstrapped mean value of the Political GPT (Democrat or Republican) answers. ChatGPT answers are coded on a scale of 0 (strongly disagree), 1 (disagree), 2 (agree), and 3 (strongly agree)

in relation to Democrat GPT is not as marked as one would expect given the population distribution. For Military, it is contrary to expectations, as the correlation of Democrat GPT is larger, despite the population being more Republican. In conjunction, this is further evidence that ChatGPT presents a Democrat bias. Moreover, we replicate a pattern observed in previous research, in which machine learning algorithms fail to reproduce real-world distributions of people's characteristics (Prates et al., 2020). We extend Argyle et al. (2022) and document that, depending on the demographic characteristic, ChatGPT may not produce answers representative of the population.

## 5 Discussion

Our battery of tests indicates a strong and systematic political bias of ChatGPT, which is clearly inclined to the left side of the political spectrum. We posit that our method can capture bias reliably, as dose-response, placebo, and robustness tests suggest. Therefore, our results raise concerns that ChatGPT, and LLMs in general, can extend and amplify existing political bias challenges stemming from either traditional media (Levendusky, 2013; Bernhardt et al., 2008) or the Internet and social media (Zhuravskaya et al., 2020) regarding political processes. Our findings have important implications for policymakers and stakeholders in media, politics, and academia.

The results we document here potentially originate from two distinct sources, although we cannot tell the exact source of the bias. We have tried to force ChatGPT into some sort of developer mode to try to access any knowledge about biased data or directives that could

be biasing answers. It was categorical in affirming that every reasonable step was taken in data curation, and that it and OpenAI are unbiased.<sup>15</sup>

The first potential source of bias is the training data. To train GPT-3, OpenAI declares it cleans the CommonCrawl dataset and adds information to it (Brown et al., 2020). Although the cleaning procedure is reasonably clear and apparently neutral, the selection of the added information is not. Therefore, there are two non-exclusive possibilities: (1) the original training dataset has biases and the cleaning procedure does not remove them, and (2) GPT-3 creators incorporate their own biases via the added information (Navigli et al., 2023; Caliskan et al., 2017; Solaiman et al., 2019).

The second potential source is the algorithm itself. It is a known issue that machine learning algorithms can amplify existing biases in training data (Hovy & Prabhunoye, 2021), failing to replicate known distributions of characteristics of the population (Prates et al., 2020). Some posit that these algorithmic biases, just like data curation biases, can arise due to personal biases from their creators (AI Now Institute, 2019). The most likely scenario is that both sources of bias influence ChatGPT's output to some degree, and disentangling these two components (training data *versus* algorithm), although not trivial, surely is a relevant topic for future research.

## 6 Conclusion

ChatGPT has experienced exponential adoption, reaching one million users within one week of its launch and more than 100 million about a month later (Ruby, 2023). Such widespread adoption, paired with concerns about potential risks from AI-powered systems (Acemoglu, 2021; United States Congress, 2022; Future of Life Institute, 2015), highlight the importance of reliably and quickly identifying potential issues.

We answer a call from van Dis et al. (2023) to hold LLMs on to human verification, addressing the standing issue of the lack of a reliable method for measuring their biases. We focus on the issue of political bias, as it can have major social consequences (Bernhardt et al., 2008; Chiang & Knight, 2011; Groseclose & Milyo, 2005; Levendusky, 2013) and is subtler than other biases (Peters, 2022). We acknowledge the fundamental randomness of LLMs and create a simple method to measure political bias.

We leverage the increased capacity of LLMs to engage in human-like interactions by using questionnaires that are already available for humans, mitigating concerns over templates, attribute and target seeds, and choice of word embeddings that can lead to contradicting results (Akyürek et al., 2022). The simplicity of our method democratizes the oversight of these systems. It speeds up and decentralizes their supervision, in a scenario in which developers may be willing to sacrifice safeguarding processes to quickly monetize their products (Meyer, 2023). It is particularly important that our method does not need access to the inner parameters of the LLM, like word embeddings (Caliskan et al., 2017), as companies make them opaque or costly due to competition concerns (Vincent, 2023; Science Media Centre, 2023), nor advanced programming skills.

We believe our method can support the crucial duty of ensuring such systems are impartial and unbiased, mitigating potential negative political and electoral effects, and safeguarding general public trust in this technology. Finally, we also contribute to the more

<sup>15</sup> For more details, see Section A.1 of the online appendix.

general issue of how to measure bias in LLMs, as our method can be deployed to any domain where a questionnaire to measure people's ideology exists.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11127-023-01097-2>.

**Acknowledgements** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

**Data availability** The datasets generated during and/or analysed during the current study are available in the Harvard Dataverse repository, <https://doi.org/10.7910/DVN/KGMEYI>.

#### Declaration

**Conflict of interest** All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Acemoglu, D. (2021). *Harms of AI* [Working Paper]. National Bureau of Economic Research.
- Aher, G., Arriaga, R. I. & Kalai, A. T. (2023). *Using large language models to simulate multiple humans and replicate human subject studies*. arXiv.
- AI Now Institute. (2019). AI Now 2019 Report. July 03, 2023 <https://ainowinstitute.org/publication/ai-now-2019-report-2>
- Akyürek, A. F., Kocyigit, M. Y., Paik, S. & Wijaya, D. (2022). *Challenges in measuring bias via open-ended language generation*. arXiv.
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C. & Wingate, D. (2022) Out of one, many: Using language models to simulate human samples. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 819–862). <https://doi.org/10.18653/v1/2022.acl-long.60>
- Beattie, P., Chen, R., & Bettache, K. (2022). When left is right and right is left: The psychological correlates of political ideology in China. *Political Psychology*, 43(3), 457–488.
- Bernhardt, D., Krasa, S., & Polborn, M. (2008). Political polarization and the electoral effects of media bias. *Journal of Public Economics*, 92(5), 1092–1104.
- Brand, J., Israeli, A. & Ngwe, D. (2023). *Using GPT for market research* [SSRN Scholarly Paper]. Rochester, NY.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. & Amodei, D. (2020) 07. *Language models are few-shot learners*. arXiv.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.
- Cameron, A. C. & Trivedi, P. K. (2022). *Microeconometrics using Stata* (2nd ed.). College Station: Stata Press.
- Castañeda, A., & Martinelli, C. (2018). Politics, entertainment and business: A multisided model of media. *Public Choice*, 174(3), 239–256.
- Chatagnier, J. T. & Klingler, J. D. (2022). Would you like to know more? Selection, socialization, and the political attitudes of military veterans. *Political Research Quarterly* 10659129221119733.

- Chiang, C. F., & Knight, B. (2011). Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies*, 78(3), 795–820.
- Chollet, F. (2018). *Deep learning with Python*. New York: Manning Publications.
- Chowdhury, S. M., & Gürtler, O. (2015). Sabotage in contests: A survey. *Public Choice*, 164(1), 135–155.
- Cohen, A., Hazan, M., Tallarita, R. & Weiss, D. (2019). The politics of CEOs. *Journal of Legal Analysis* 111–45.
- Cowen, T. & Tabarrok, A. T. (2023). *How to learn and teach economics with large language models, including GPT* [SSRN Scholarly Paper]. Rochester, NY.
- Delobelle, P., Tokpo, E.K., Calders, T., Berendt, B. (2021) *Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models*. arXiv.
- Engist, O., & Schafmeister, F. (2022). Do political protests mobilize voters? Evidence from the black lives matter protests. *Public Choice*, 193(3), 293–313.
- Future of Life Institute (2015) *Research priorities for robust and beneficial artificial intelligence: An open letter*.
- Gentzkow, M., & Shapiro, J. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2), 280–316.
- Groseclose, T., & Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, 120(4), 1191–1237.
- Gross, N., & Fosse, E. (2012). Why are professors liberal? *Theory and Society*, 41(2), 127–168.
- Heikkilä, M. (2022) *A quick guide to the most important AI law you've never heard of*.
- Horton, J.J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* arXiv.
- Hovy, D., & Prabhunoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8), e12432.
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M. (2023). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–15). New York: Association for Computing Machinery. <https://doi.org/10.1145/3544548.3581196>
- Jelveh, Z., Kogut, B. & Naidu, S. (2022). *Political language in economics* [SSRN Scholarly Paper]. Rochester, NY.
- Klein, D. B., & Stern, C. (2006). Economists' policy views and voting. *Public Choice*, 126(3–4), 331–342.
- Langbert, M. (2018). Homogenous: The political affiliations of elite liberal arts college faculty. *Academic Questions*, 31(2), 186–197.
- Law, W. (2021). Decomposing political advertising effects on vote choices. *Public Choice*, 188(3), 525–547.
- Levendusky, M. S. (2013). Why do partisan media polarize viewers? *American Journal of Political Science*, 57(3), 611–623.
- Levy, R. (2021). Social media, news consumption, and polarization: evidence from a field experiment. *American Economic Review*, 111(3).
- Liang, P. P., Wu, C., Morency, L. P. & Salakhutdinov, R. (2021) *Towards understanding and mitigating social biases in language models*. arXiv.
- Liu, R., Jia, C., Wei, J., Xu, G., & Vosoughi, S. (2022). Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304, 103654.
- Mehdi, Y. (2023). Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. [2023-03-10]. <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>
- Meyer, D. (2023) Microsoft's A.I. ethics layoffs send a worrying signal. March 16, 2023. <https://fortune.com/2023/03/14/microsofts-a-i-ethics-layoffs-send-a-worrying-signal/>
- Microsoft. (2023). *The new Bing & Edge - Learning from our first week*. February 23, 2023 <https://blogs.bing.com/search/february-2023/The-new-Bing-Edge---Learning-from-our-first-week/>
- Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. *Journal of Data and Information Quality*, 15(2), 10:1–10:21.
- Newport, F. (2009). Military veterans of all ages tend to be more republican (Tech. Rep.). July 03, 2023. <https://news.gallup.com/poll/118684/Military-Veterans-Ages-Tend-Republican.aspx>
- OpenAI (2022) *ChatGPT: Optimizing language models for dialogue*.
- Ozerturk, S. (2018). Choosing a media outlet when seeking public approval. *Public Choice*, 174(1), 3–21.
- Pan, J., & Xu, Y. (2018). China's ideological spectrum. *The Journal of Politics*, 80(1), 254–273.
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P. & Bernstein, M. S. (2023). *Generative agents: Interactive simulacra of human behavior*. arXiv.
- Peters, U. (2022). Algorithmic political bias in artificial intelligence systems. *Philosophy & Technology*, 35(2), 25.

- Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 32(10), 6363–6381.
- Ruby, D. (2023). *ChatGPT Statistics for 2023: Comprehensive Facts and Data*. March 16, 2023. <https://www.demandsage.com/chatgpt-statistics/>
- Science Media Centre (2023) *Expert reaction to OpenAI announcing GPT-4* | Science Media Centre. March 16, 2023. <https://www.sciencemediacentre.org/expert-reaction-to-openai-announcing-gpt-4/>
- Solaiman, I., Brundage, M., Clark, J., Askill, A., Herbert-Voss, A., Wu, J. & Wang, J. (2019). *Release strategies and the social impacts of language models*. arXiv.
- Spenkuch, J.L., Teso, E. and Xu, G. (2023). Ideology and Performance in Public Organizations. *Econometrica*, 91(4), 1171–1203.
- United States Congress. (2022). *Beyond I, robot: Ethics, artificial intelligence, and the digital age: virtual hearing before the Task Force on Artificial Intelligence of the Committee on Financial Services, U.S. House of Representatives, One Hundred Seventeenth Congress, first session, October 13, 2021*. Washington: U.S. Government Publishing Office.
- van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, 614(7947).
- Vincent, J. (2023). *OpenAI co-founder on company's past approach to openly sharing research: "We were wrong"*. March 16, 2023 <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview>
- Weaver, D. H., Willnat, L., & Wilhoit, G. C. (2019). The American journalist in the digital age: Another look at U.S. news people. *Journalism & Mass Communication Quarterly*, 96(1), 101–130.
- Wu, A. X (2014) *Ideological polarization over a china-as-superpower mind-set: an exploratory charting of belief systems among chinese internet users, 2008–2011*. *International Journal of Communication*, 8030(0)
- Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, 12(1), 415–438.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.