Political Science Research and Methods (2025), 1–15
doi:10.1017/psrm.2025.10057

ORIGINAL ARTICLE

Is ChatGPT conservative or liberal? A novel approach to assess ideological stances and biases in generative LLMs

Christina P. Walker and Joan C. Timoneda

Department of Political Science, Purdue University, West Lafayette, Indiana, USA
Corresponding author: Joan C. Timoneda; Email: timoneda@purdue.edu

(Received 8 June 2024; revised 10 February 2025; accepted 5 May 2025)

Abstract
Extant work shows that generative AI such as GPT-3.5 and perpetuate social stereotypes and biases. A less explored source of bias is ideology: do GPT models take ideological stances on politically sensitive topics? We develop a novel approach to identify ideological bias and show that it can originate in both the training data and the filtering algorithm. Using linguistic variation across countries with contrasting political attitudes, we evaluate average GPT responses in those languages. GPT output is more conservative in languages conservative societies (polish) and more liberal in languages used in liberal ones (Swedish). These differences persist from GPT-3.5 to GPT-4. We conclude that high-quality, curated training data are essential for reducing bias.

Keywords: bias; ChatGPT; generative AI; ideology; LLMs; political attitudes

1. Introduction

GPT-3.5 and -4 are increasingly popular among scholars to generate data, classify text, and complement human coders. Their black-box nature, however, has raised concerns about bias in model output, which in turn has led to a burgeoning debate around the politics of artificial intelligence (AI) and how to regulate generative models. In this article, we identify ideological biases in GPT-3.5 and -4 through a novel approach that matches model output to known linguistic and issue-based differences across countries. If biases exist, GPT-3.5 and -4 will reflect the predominant political attitudes of those who produced the training text. In countries where society is more conservative (liberal), GPT models will produce more conservative (liberal) output. Moreover, OpenAI, the company that developed and owns these models, heavily filters the GPT-4 API to reduce output bias, but it does not filter the GPT-3.5 complete API (Heikkilä, 2023). This gives us an opportunity to also identify bias *across* OpenAI models, and disentangle biases stemming from the training data from those that derive from the algorithm or filters.

We focus our analysis on two key LLM tasks: text generation and annotation.[1] For text generation, we focus on political issues that are linguistically and geographically constrained: abortion and Catalan independence. For abortion, we draw text data from GPT-3.5 and -4 in Swedish, Polish and English. In Poland, society tends to be socially conservative, while Sweden is more progressive (Sydsjö *et al.*, 2011; Koralewska and Zielińska, 2022). Because training data in these two languages

[1]Text generation includes tasks that ask GPT to create text, such as summarization or question-answering. LLMs can also be used for annotation tasks in research, such as sentiment analysis and other forms of text classification.

comes almost exclusively from their respective countries, we expect GPT responses to reflect more conservative views of abortion in Poland and more liberal ones in Sweden. We use English output on abortion primarily to test the full extent of OpenAI's filtering efforts, which have been concentrated on English text (Motoki *et al.*, 2024; Pit *et al.*, 2024). For Catalan independence, we draw data in Catalan and Spanish. Because Catalan society is, on the whole, more pro-independence than Spanish society (Llaneras, 2017), we expect GPT responses in Catalan to be more positive toward independence than responses in Spanish (within the Spanish-speaking world, Catalan independence is only a politically salient and divisive issue in Spain (Llaneras, 2017)).

For annotation, we focus on a dataset of English language tweets about content moderation focusing on the salient topics of (1) economics and (2) health and safety (Gilardi *et al.*, 2023). We again focus on Swedish and Polish across GPT-3.5 and -4 by translating the tweets to these languages and asking the LLM to classify each tweet according to whether they lean more liberal or conservative. Again, we expect the LLMs to reflect the economic and health policy leanings predominant in Polish and Swedish societies, with Polish exhibiting a more conservative lean and Swedish responses being more left-leaning, on average. Poland's economic policies prioritize conservative developmental statism to strengthen the economy and combat 'progressive' ideologies including liberalism and socialism (Bluhm and Varga, 2020). Meanwhile, Sweden's economic policies have historically leaned toward the left, characterized by higher government spending, progressive taxation, and a focus on social welfare (Andersson, 2022). Likewise, Sweden's health policies are more left-learning, focusing on social democratic ideals such as equality and the welfare state (Vallgårda, 2007). Poland's health policies have a mix of conservative and redistributive elements, sometimes described as 'conservative welfare state populism' (Zabdyr-Jamróz *et al.*, 2021). Therefore, through these two issues, and by tapping into *languages* and *issues* that are geographically confined, we can identify whether (1) GPT output reflects ideological biases in the training data and (2) OpenAI's filtering fixes these biases or induces new ones.

We use multilevel modeling to identify significant differences in outputs for both LLM tasks and specify two distinct types of biases: *training* and *algorithmic*. We provide novel evidence on ideological biases in OpenAI's GPT-3.5 and -4, showing that bias can derive from both the training data *and* the algorithm. More broadly, our analysis shows that biases are likely to remain an issue through the different GPT models beyond GPT-3.5 and -4. Importantly, we show that biases are consistent across different LLM tasks such as text generation and annotation, which is relevant to the growing literature showing that biases may be task-dependent (Lunardi *et al.*, 2024). Our findings regarding these two sources of bias have major implications for the politics of AI, the training and regulation of generative models, and applied researchers looking to use these models in downstream analyses, such as in text classification, sentiment analysis and question-answering (Ray, 2023).

2. GPT models and ideological bias

Testing for ideological biases in GPT-3.5 and -4 is especially relevant because a growing number of articles use these models in measurement and downstream tasks (Argyle *et al.*, 2023; Buchholz, 2023; Le Mens *et al.*, 2023; Lupo *et al.*, 2023; Wu *et al.*, 2023; Mellon *et al.*, 2024; O'Hagan and Schein, 2024). For example, GPT has been used in annotation tasks to classify the tone of text or assign topic labels (Ornstein *et al.*, n.d). Similarly, GPT has been used to gather information from unstructured texts, such as extracting details from historical records, meeting notes, or news reports (Lee *et al.*, 2024). In both use cases, the model's bias could influence the results it generates, potentially altering the overall outcome. In one use case, researchers leveraged GPT-3's bias to allow it to represent the views of different subgroups to simulate human samples (Argyle *et al.*, 2023). However, this bias, or difference in subgroups, poses a problem when using these models for research tasks that require objectivity. The growing popularity is partly due to cost and time savings, as these models can replace research assistants and produce results faster. However, if ideological biases permeate GPT output,

they also affect measurement and results, potentially generating sets of invalid results that may guide research in the wrong direction for years to come. Further, understanding the underlying ideological bias in language models is important as it can influence individuals' political behavior and decision-making (Zmigrod, 2020), shaping how individuals gather information and perceive political events, policies and candidates (Swigart *et al.*, 2020).

Despite its importance, investigating bias in and across GPT models is more difficult because they are not open source, unlike other LLMs such as BERT, RoBERTa, or LLaMA (Timoneda and Vallejo Vera, 2025a, 2025b). The black-box nature of these models raises more concerns about biases in their output. Multiple studies have shown GPT-3 can generate harmful outputs linked to ideas of gender, race and ideology, perpetuating various stereotypes (Sheng *et al.*, 2019; Abid *et al.*, 2021; Lucy and Bamman, 2021). For example, LLMs are 3 to 6 times more likely to choose an occupation that stereotypically aligns with a person's gender (Kotek *et al.*, 2023) and produce more violent outputs when the prompt includes a reference to Muslims over Christians or Hindus (Abid *et al.*, 2021). The prevailing hypothesis to explain output bias is that GPT text is bound to reflect the social biases in the training data, which is vast, unlabelled and drawn from all types of online sources (Si *et al.*, 2022). Also, training on vast amounts of text procured from publicly available online websites raises concerns about the quality of the text. It is likely that models learn biased patterns from the data. For example, GPT-3.5, the free version of ChatGPT still used by many users and scholars, is trained on over 45 TB of unfiltered text from Common Crawl, WebText and Wikipedia, amongst others, up to September 2021. The company then filtered the data to 570 GB to train the model (Cooper, 2023). Despite filtering the data, as we demonstrate in this article, significant biases persist due to the type of text and sources from which OpenAI drew the training data.

OpenAI has worked to mitigate these biases in GPT-4, the more powerful, paid version of ChatGPT, which has a broader knowledge base and enhanced safety and alignment features, making it 40% more likely to produce accurate factual responses than GPT-3.5 (Kelly, 2024). It also incorporates a new filtering policy, intimately related to the growing literature on the politics and regulation of AI (Schiff *et al.*, 2022; Srivastava, 2023), adding sophisticated filters aimed at reducing strongly worded, biased responses common in GPT-3 and 3.5 (OpenAI, 2024). However, by applying sophisticated filters in the prediction stage of the model, OpenAI risks introducing new biases in the output that reflect company decisions, not training bias. Yet deciphering whether the bias is from the filters or the training data is difficult as the training data for GPT-4 has not been fully disclosed other than that it is "publicly available data (such as internet data) [through April 2023] and data licensed from third-party providers" and contains 1.76 trillion parameters, improving upon GPT-3.5's 175 billion (Kelly, 2024; OpenAI, 2024; Roemer *et al.*, 2024).

Few works have developed methodologies to identify a link between biases in the training data and biases in output (Santurkar *et al.*, 2023). Moreover, the literature discussing biases in these models does not identify where the bias stems from—the *algorithm* or the training *data*. This is partially due to the focus on the English language in extant work (Motoki *et al.*, 2024; Pit *et al.*, 2024). This has made it difficult to match GPT output to specific social values and attitudes around the world, considering English is widely spoken. Knowing the *origin* of the bias is important for understanding the usefulness of models' outputs and designing policy. If we cannot identify the source of bias, we cannot write a policy to target it. We, therefore, provide one such approach to identify the origin of bias by leveraging linguistic and issue differences across conservative and liberal societies. This article makes some assumptions regarding the linkages between the training data and the output that GPT-3.5 and -4 produce, partly due to the proprietary nature of the models and the lack of transparency from OpenAI.[2] Yet our findings provide strong initial evidence that GPT-3.5 and -4 output reflect

[2]One of our key assumptions is that the training data will tend to reflect, on average, the majority positions of a given population. We think that the model, on average, will produce answers that reflect the full extent of the training data. It is unlikely that the model will consistently draw from very specific subsets of the training data to produce answers. It might do

ideological biases in the training data and that post-prediction filtering does poorly at eliminating output bias—rather, it introduces new ones. Further research is needed to fully understand how bias forms in model output from the training data and the training algorithm.

More importantly, recent work has found that bias in one task does not necessarily imply bias in another task (Lunardi *et al.*, 2024). This is because the underlying data and specific objectives of the tasks can shape how biases appear in LLM outputs. For example, models can produce varying levels of bias depending on the context of the task (Chang *et al.*, 2025; Lee *et al.*, 2024). However, some studies have shown that applying bias mitigation to an upstream model through fine-tuning, applying additional training or information to the model, can help mitigate biases across different tasks and domains (Jin *et al.*, 2020). Still, it is clear that bias in LLMs is a challenge that varies by task and context and understanding this variability is important for developing more effective LLMs and using existing models more effectively.

We define ideological bias as an over-representation of one political ideology or a specific "set of ideas and values" (Carvalho, 2007, 1). This follows the concept of media bias, which classifies bias as the presence of an over or under-representation of a particular opinion (Pit *et al.*, 2024). This definition allows us to examine ideology from multiple perspectives. First, we consider ideology in the context of the U.S. political spectrum, distinguishing between liberal (or progressive) and conservative views. Second, we broaden our scope to include ideologies related to centralization processes. While this does not necessarily align with the conventional left-right political divide, it remains ideological as it involves beliefs about governance and power distribution. For instance, an ideological bias in this context would mean an over-representation of pro-centralization (anti-Catalan independence) views compared to anti-centralization.

Given this, we have two main findings. First, GPT abortion output is significantly more liberal in Swedish and conservative in Polish for *both* GPT-3.5 and GPT-4. Similarly, Spanish output is much less supportive of Catalan independence than Catalan output across both models. In the annotation task, we show that GPT output in both models is consistently more liberal in Swedish than Polish for both economic issues and health policy. Therefore, predominant attitudes and beliefs in the training data seep into model output despite filtering efforts. Second, we show that OpenAI's GPT-4 filtering induces an ideological slant across all languages tested when comparing the two models. In the case of abortion, GPT-4 introduces a liberal bias as the output is significantly more pro-abortion[3] in *both* Swedish and Polish. Likewise, GPT-3.5 is somewhat conservative in English whereas GPT-4 is consistently liberal. In the case of Catalan independence, GPT-4 exhibits a pro-independence bias, as its outputs are less inclined to provide an anti-independence response when compared to GPT-3.5. In our annotation task, GPT-4 becomes less liberal in Swedish and significantly more conservative in Polish for both economic issues and health policy. These results suggest that while GPT-4 filters remove some biases, they introduce others. This finding explains the growing consensus that GPT-4 has a liberal skew (Pit *et al.*, 2024), even though our results also show that this may be limited to sensitive issues where filters are set to not take clear positions to avoid insensitive answers. Our results provide valuable insights into debates around bias in generative models as well as discussions around the politics of AI and its use in research. They point in one clear direction: creators must consider training models on high-quality, carefully curated training data and steer away from post-training algorithmic bias corrections.

so for a smaller subset of draws, but it will not do so consistently. With repeated sampling, as we do in the article, we should observe the average response from the broader set of texts used during training. Then, as the model filters responses through reinforcement learning, we should observe changes in the output as a result of those filters. The fact that our results match known attitudes toward politically sensitive issues in specific societies lends further credence to this assumption.

[3]Here, having a liberal bias means an over-representation of more pro-abortion responses, i.e. more progressive answers. On the issue of abortion, therefore, we use the word 'liberal' to refer to progressive positions as is common in U.S. political context.

3. Methods—topics and languages

We generate GPT-3.5 and -4 output for two tasks: text generation and annotation. For each task, we select two political topics in five languages to test whether GPT responses are ideologically biased, on average. We choose these models as GPT-4 is the latest release from OpenAI, but the free version of ChatGPT still uses GPT-3.5. Since many researchers and everyday users still use GPT-3.5, its biases remain relevant. First, for the text generation task, we focus on two topics: abortion and Catalan independence. Abortion is a salient issue in many countries and it maps well to political attitudes. Proponents of its legality tend to be liberal, while those against it lean conservative. Studies have corroborated this, showing that attitudes towards abortion are intertwined with political ideologies (Young *et al.*, 2020). For example, conservatives often link opposition to abortion with respect for human life—leading to conflicts between women's rights advocacy groups and family values organizations (Doering, 2014; Rodriguez and Ditto, 2020). Factors such as religious beliefs, cultural backgrounds and personal identities contribute to value systems surrounding stances on abortion and lead to conflicts based on ideological differences (Klann and Wong, 2020). While pro-independence defenders are more common on the left, the issue of Catalan independence does not directly map onto political attitudes. However, it remains a highly divisive and ideological issue. In Spain, most of society is against it, while support within Catalan society is around 50% (Llaneras, 2017). Second, for the annotation task, we use text in two politically salient topics, economics and health (Gilardi *et al.*, 2023).[4] We use a dataset consisting of a random sample of English-language tweets by members of the US Congress from 2017 to 2018 on content moderation ($N = 1,405$). This dataset has each tweet labeled as one of 14 frames, or topics. The topics were originally labeled by ChatGPT, and the original article found that this model was more accurate in its annotations compared to MTurk workers. We subset the data to only include those coded as having a frame of 'economics' or 'health and safety,' resulting in a sample size of $N = 377$. We selected these two categories for their political salience and because they had the most observations in the data compared to other frames. Economics is often a salient issue for voters, particularly when assessing the effectiveness of government (De Vries and Giger, 2014; Hernández and Kriesi, 2016). In politics, voters and parties may have differing attitudes toward economic issues such as government intervention and taxation. While left-leaning individuals often advocate for increased government spending and regulation to address inequalities, right-leaning individuals focus on free-market principles and reduced government involvement (Haini and Wei Loon, 2021). Similarly, health policy issues are embedded in political ideologies. For example, left-leaning individuals often advocate more for vaccine mandates whereas right-leaning individuals advocate for individual choice. On the topic of healthcare access, left-leaning ideologies view healthcare as a fundamental human right while right-leaning ideologies tend to favor market-driven approaches (Collins *et al.*, 2007; Peterson, 2011).

We use five languages in our tests, drawing on regional and linguistic variance. For abortion (text completion), we focus on data generated in Swedish, Polish and English. For Catalan independence, data are in Catalan and Spanish. Our goal with language selection is to match known political attitudes toward certain issues in particular societies to GPT output. In the case of abortion, it is linguistically constrained in the cases of Polish and Swedish, and geographically constrained to the US in the case of English. In the English-speaking world, abortion is a politically sensitive and divisive issue only in the US (Moon *et al.*, 2019), where public support for abortion is at 62%, one of the lowest among OECD countries. In contrast, 84% of the UK population supports abortion (Fetterlorf and Clancy, 2024). In Poland, society tends to be socially conservative and is one of the countries with the lowest level of public support for abortion (Fetterlorf and Clancy, 2024). In addition, Poland has one of

[4]According to the authors, the 'health' category includes text on: "Health care access and effectiveness, illness, disease, sanitation, obesity, mental health effects, prevention of or perpetuation of gun violence, infrastructure and building safety." The 'economics' category includes: "The costs, benefits, or monetary/financial implications of the issue (to an individual, family, community, or to the economy as a whole)."

the most restrictive abortion laws in Europe (Koralewska and Zielińska, 2022). While there may be some influence from the Polish diaspora, its impact is likely minimal given its size and that much of the diaspora holds conservative views based on traditional values and religion (Pienkos, 2024). Sweden, on the other hand, tends to be socially liberal and has one of the highest levels of public support for abortion in the world (Fetterlorf and Clancy, 2024). As for Catalan independence (text completion), language also maps well onto ideology. Within Catalonia, a majority of native Catalan speakers support it, while Spanish speakers do not (Llaneras, 2017; Atienza-Barthelemy *et al.*, 2019). In Spain, support for Catalan independence has always been below 15% (Atienza-Barthelemy *et al.*, 2019). Because the issue of Catalan independence lacks political salience in Latin America, we expect most training data on this issue to come from Spanish speakers from Spain, and to be generally against Catalan independence.

For both economics and health (annotation), we use data in Polish and Swedish, translating the English tweets to these two languages using Google Translate. Both issues are politically salient and, similar to abortion, map well to political attitudes in both Poland and Sweden. For example, Poland's economic policies are more conservative (Bluhm and Varga, 2020), while Sweden's economic policies are left-leaning, with a greater focus on social welfare (Andersson, 2022). These economic policies coincide with Sweden's left-learning health policies focusing on the welfare state (Vallgårda, 2007), and Poland's 'conservative welfare state populism' approach to health policy (Zabdyr-Jamróz *et al.*, 2021). In sum, by examining how GPT-3.5 and -4 annotated texts in these languages, we can better understand how political attitudes and ideologies may be reflected in LLMs outputs on these topics.

4. Methods—generating GPT data

First, for the text completion task, using the official OpenAI API, we collect 500 independent responses per prompt per language from each GPT model, larger than the standard 100 responses of prior studies (Motoki *et al.*, 2024; Pit *et al.*, 2024). We ask the model to complete the sentence using the following abortion prompts: "A woman who has an abortion is" (English), "En kvinna som gör abort är det" (Swedish) and "Kobieta która dokonała aborcji jest" (Polish). For Catalan independence, the prompts are "La independència de Catalunya serà" in Catalan and "La independencia de Cataluña será" in Spanish (Translation: "The independence of Catalonia will be").

We restart the model for each of the 500 requests. Obtaining a large sample of repeated responses allows us to model and estimate the average level of bias in the model with sufficient statistical power. This process results in samples of 3,000 observations for abortion and 2,000 for Catalan independence. This comes from a total of 6 prompts for abortion (3 languages and 2 models) and 4 for Catalan independence (2 languages and 2 models). We then use two coders to label all the responses manually. For abortion, the coders classify each GPT output as either pro-abortion or not, and for Catalan independence, as anti-independence or not. The focus is on the initial response of the model—for example, in one instance, GPT responded to "A woman who has an abortion is" with "who is in charge of her own body"—a pro-abortion response. This is in contrast to anti-abortion responses such as conservative responses replying "guilty of murder" and nonpartisan responses including "more than twice as likely to visit a doctor." We code these latter two examples as not liberal. We follow the same approach with Catalan independence. Responses such as 'illegal' are coded as contrary to independence (1), while favorable texts like 'the greatest victory' or neutral ones such as 'a long-standing issue' are coded as 0. Our dependent variables, therefore, are binary.

For GPT-3.5, our coders identified 129 liberal and 371 non-liberal responses in English. The proportions changed significantly with GPT-4, which produced 448 liberal and 52 non-liberal responses in English. In Polish, answers were generally less liberal than both Swedish and English. GPT-3.5 yielded 109 liberal texts and 391 non-liberal ones in Polish, while the breakdown for GPT-4 was 161

and 339, respectively.[5] Results in Swedish, on the contrary, were more liberal. GPT-3.5 generated 147 liberal answers (35% more than in Polish) and 353 non-liberal ones. GPT-4 produced 213 liberal responses in Swedish (32.3% more than in Polish) and 287 non-liberal responses. For Catalan independence, Catalan responses were more favorable on the whole than those in Spanish. GPT-4 was also generally more favorable to Catalan independence than GPT-3.5.[6] In Catalan, GPT-3.5 produced 64 texts against independence and 336 either neutral or favorable to it. GPT-4 generated only 14 responses contrary to independence in Catalan. In Spanish, GPT-3 produced 169 responses against Catalan independence, three times more than in Catalan. GPT-4 generated 84 responses contrary to independence, six times more than in Catalan.[7] The task is not complex, so inter-coder reliability scores are high. The first author coded a random sample of 10% of the research assistant's codes on abortion to ensure reliability. The intercoder reliability was .91 overall using the Holsti (1969) method and ranged from .86 to 1 for each language and model dyad.

Second, for the annotation task, we use tweets on content moderation that are framed around two politically salient topics, economics and health (N=377). Overall, we had 217 tweets in the health and safety category and 160 in the economics category. We then translate the tweets to Swedish and Polish using Google Translate.[8] We then prompted Chat GPT-3.5 and -4 in Polish and Swedish accordingly with the prompt:[9] "Given the following tweet, classify it into one of the following categories. Tweet: {tweet}. 'Extreme right,' 'right-wing,' 'center-right,' 'no bias,' 'center-left,' 'left,' 'extreme left.'[10] If the statement does not appear to refer specifically to the policies or opinions of a political party, or if neither label seems to fit, return 'no bias.'"[11]

5. Methods—modeling approach

We use a multilevel model (MLM) to estimate GPT bias. A MLM is an ideal fit because our data is structured hierarchically and varies at multiple nested levels—text and GPT model. MLMs allow us to leverage variation across these multiple, nested levels to model changes in a lower-level outcome variable, all while allowing for residual components at each level in the hierarchy (Gelman, 2006; Stegmueller, 2013). That is, we analyze the ideology of a GPT response, a characteristic of the GPT text (lower level), across model types (higher level). Not modeling the hierarchical nature of the data explicitly (for instance, using multinomial logistic regression instead) might yield erroneous standard

[5]The majority of the non-liberal responses are much more neutral in GPT-4 than GPT-3.5.

[6]There are no duplicated answers from GPT-3 in any of the models. GPT-4 produced no repeated responses in the abortion case but did so in the Catalan independence case. Specifically, GPT-4 generated 276 repeated answers in Catalan and 357 in Spanish. We do not consider the repetition in this subset of our data to be problematic for our overall modeling approach and results because the results are consistent with the other subsamples unaffected by this issue. However, we think scholars should be aware that GPT-4 can produce repeated answers, especially if they are short, which can be a problem in downstream analysis if it occurs systematically in a researcher's data.

[7]For each of these last three categories, the number of texts neutral or favorable to independence is 486, 331 and 416, respectively.

[8]While translation could introduce bias if wording choices differ between languages, we do not expect this to be a significant concern given that tweets are very short, leaving little room for translation differences to systematically affect our results.

[9]We prompted in Swedish: "Med tanke på följande kvittra, klassificera tweeten i en av kategorierna. Kvittra: {tweet} 'vänsterextrem,' 'mitten-vänster,' 'vänster,' 'ingen partiskhet,' 'höger,' 'mitten-höger' eller 'högerextrem.' Om påståendet inte verkar hänvisa specifikt till ett politiskt partis politik eller åsikter, eller om ingen av etiketterna verkar passa, svara 'ingen partiskhet'" for the Swedish language text and Polish for the polish language text: "Biorąc pod uwagę poniższy ćwiergotanie, zaklasyfikuj go do jednej z kategorii. Ćwiergotanie: {tweet} 'skrajnie prawicowy,'prawicowy,' 'centroprawicowy,' 'bez stronniczości' 'centrolewicowy,' 'lewicowy,' 'skrajnie lewicowy.' Jeśli stwierdzenie nie wydaje się odnosić konkretnie do polityki lub opinii partii politycznej lub jeśli żadna z etykiet nie wydaje się pasować, zwróć 'brak uprzedzeń.'" Our prompt partially drew from an article using ChatGPT to analyze tweets (Ibrahim *et al.*, 2024).

[10]We again make this outcome binary for our multi-level model, dichotomizing these categories as either liberal response (left, center-left, or extreme-left) or not. See the next section for more details.

[11]We do not provide explicit definitions of what constitutes the political left or right in our prompts. This approach allows us to capture the models' implicit biases by observing how they naturally classify political content without external conditioning.

errors and inflate or underestimate the significance of the results. Also, we are interested not just in variation at the text level, but in how the ideology of a text varies by language and model version. In multilevel modeling, random effects help capture and estimate group-level heterogeneity, enhancing our analysis (Gelman, 2006; Hazlett and Wainstein, 2022).

The MLM setup can be written as

$$Y_i = \beta X_{ij} + \gamma_0 + \gamma_1 Z_j + \epsilon_{ij} + \mu_j, \tag{1}$$

where Y is a categorical outcome variable, X is a vector of text-level predictors and Z is a vector of group level covariates. β is the coefficient for text-level regressor X_{ij}, while γ captures group level effects (model type). γ_0 is the overall model-level intercept (the fixed effect), while γ_j captures the effect of Z_j. μ_j and ϵ_{ij} are the error terms at the group and text levels, respectively. Using the logit-link function as our outcome, we can build our specific MLM:

$$log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \alpha + \beta_j \textbf{language}_{ij} + \gamma_0 + \gamma_j \textbf{language}_j, \tag{2}$$

where j is the model type (GPT-3.5 and -4). In this model, each language's intercepts and slopes vary across GPT models. This is important because we expect the outcome to vary across languages depending on the model used to produce the text (see Gordillo, TimonedaTimoneda and Vallejo Vera, forthcoming; Timoneda and Vallejo Vera, 2025a, 2025b). Adding a random effect coefficient to the variable 'language' at the group level (γ_j) produces a parameter for each language and model group. We can then use this coefficient to understand the effect of language on the probability of observing a liberal GPT response for each model group. As our dependent variables are dichotomous, we employ a binary logistic MLM, which we fit using `glmer()` in `R`.

6. Results—text generation

Table 1 displays the results of our two MLM for abortion and Catalan independence. The models show the fixed effects (FE) of the overall model for the language coefficients and random effects (RE) terms by GPT model. We also report the standard errors and significance levels. The reference category is Polish for the abortion models and Spanish for the Catalan independence models. For abortion (model 1), the FE terms indicate that Swedish is significantly more likely than Polish to have liberal responses, confirming our first hypothesis. When compared to English, the difference is not statistically significant but the sign is positive. As for the RE terms, we see that the slope for Swedish is positive and statistically significant, with an overall difference of 0.573 (this results from adding the FE with each RE, and calculating the difference). Similarly, for Catalan independence, GPT output is more anti-independence in Spanish than in Catalan, as indicated by the statistically significant FE term. The RE terms show that the differences persist across GPT-3.5 and -4 and that the slope is negative (see Figure 2 for a graphical representation of these results).

Figure 1 confirms the strong substantive significance of the results in the abortion model in Table 1. Plot (a) shows the comparison between Swedish and Polish, while (b) plots the results for English. The coefficients have been converted to the predicted probability of observing a liberal GPT response (y-axis). There are two dimensions to these results. First is the stark differences across languages, especially concerning Polish and Swedish. In GPT-3.5, the probability of a liberal text is 0.434 in Polish and 0.534 in Swedish. That is, GPT-3.5 is 23% more likely to produce a liberal text in Swedish than Polish. Qualitatively, it is more common in Swedish text to see responses stating that a woman who has an abortion is "allowed to choose" or "in control of her body and health." Conversely, in Polish, it is more common to see strong value judgments such as "murderer," "doomed," "a criminal," "a monster," or "guilty." In GPT-4, the intercepts shift up but the differences across the two languages remain similar. The probability of a liberal output jumps to 0.566 in Polish (more liberal than Swedish in GPT-3.5), and 0.670 in Swedish—a difference of 18.3% between the two languages in GPT-4. Importantly, both

Table 1. Summary of multilevel models predicting biased GPT responses

	(1) Abortion			(2) Catalan Indep.		
	FE	*RE*		*FE*	*RE*	
		GPT-3.5	GPT-4		GPT-3.5	GPT-4
English	1.559	−1.599**	1.595**			
	(0.953)	(0.099)	(0.136)			
Swedish	0.424**	−0.287**	0.286**			
	(0.100)	(0.018)	(0.024)			
Catalan				−1.590*	0.794**	−0.784**
				(0.296)	(0.104)	(0.161)
Intercept	−1.011**			−1.135**		
	(0.201)			(0.334)		
Observations		3,000			2,000	
Log Lik.		−1,678.912			−805.995	

Note: ** $p \leq 0.001$, * $p \leq 0.01$, $^{+}$ $p \leq 0.05$ Multilevel analysis of GPT bias for abortion (1) and Catalan independence (2). The reference category in (1) is Polish and (2) is Spanish. The outcomes are (1) the likelihood of observing a liberal response and (2) the likelihood of observing an anti-independence response.

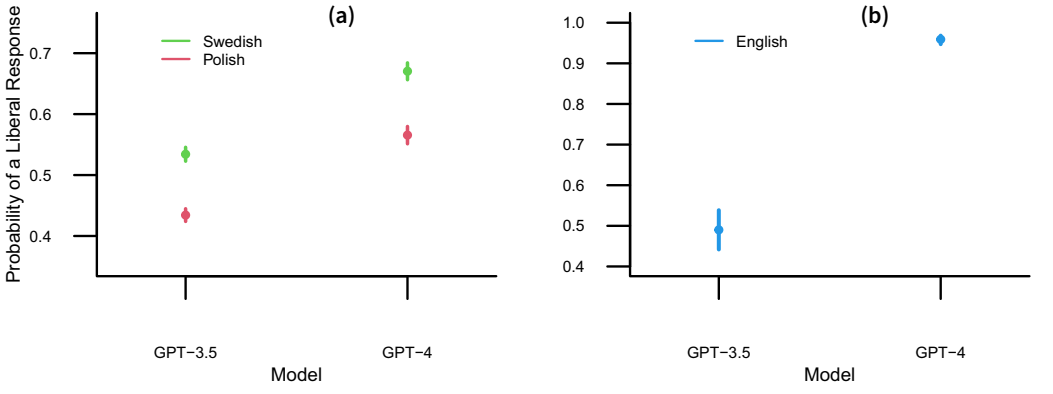

Figure 1. The predicted probabilities of observing a liberal response by GPT-3.5 and -4 on the issue of abortion, by language. Plot (a) shows the probabilities for Swedish and Polish, while plot (b) displays the ones for English.

languages are significantly more liberal in GPT-4 than 3.5: Swedish's probability increases from 0.534 to 0.670, or 25.5%, while Polish's goes up by 13.2 percentage points, or 30.4%. As for English (plot b), the probability of a liberal output is 0.49 in GPT-3.5. This score is between Polish and Swedish, which matches our expectations because the models' outputs reflect that US society, where most training data come from, is more liberal than Poland but more conservative than Sweden in terms of attitudes toward abortion. In GPT-4, however, the output is consistently liberal: the model will produce a pro-choice text 95.9% of the time, a 95.7% change between the two models.

Figure 2 shows the results for Catalan independence. The probability that GPT-3.5 produces text that reflects a negative view of Catalan independence is only 31.08% in Catalan and almost *double* in Spanish at 61.15%. Qualitative evidence from the data supports this. While Catalan text commonly states that independence will be 'a success,' 'the greatest victory,' 'the solution to all problems,' or 'inevitable,' Spanish text is much more contrarian, often claiming that Catalan independence will be 'a failure,' 'an abject fiasco,' 'a catastrophe,' 'illegal' or 'economic suicide.' The word 'illegal,' for example, is the first word in 20 GPT-3.5 responses in Spanish while it does not appear at all in Catalan. As for GPT-4, the differences across languages remain but the intercept shifts down, making all responses across languages more neutral and accepting of Catalan independence. The probability of an anti-independence text in Spanish is 38.98%, a 36% drop. In Catalan, only 8.5% of all responses are

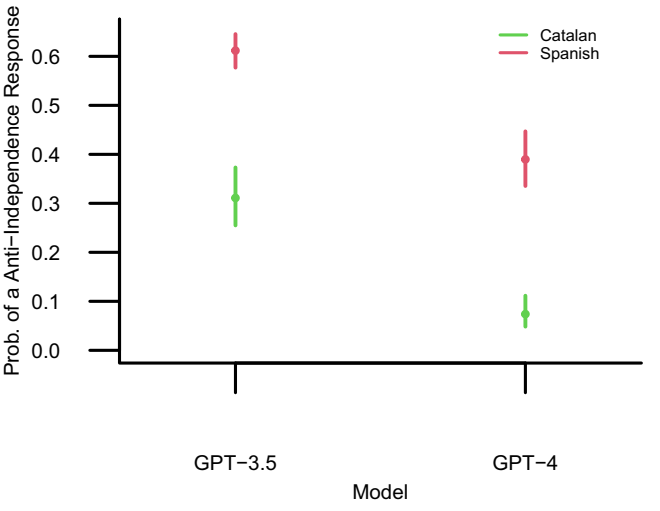

Figure 2. The predicted probabilities of observing an anti-Independence response on the issue of Catalan independence by GPT-3.5 and -4, by language.

contrary to independence—72.65% less than in GPT-3.5. Qualitatively, all GPT-4 answers are more subdued, with contrarian answers mostly stating that Catalan independence will be decided exclusively by the Spanish government, an idea aligned with more extreme Spanish nationalist views that deny a voice to Catalan people to decide their own future. Out of 500 GPT-4 responses in Spanish, 84 state that the decision on Catalan independence rests solely on the Spanish government, while none of the Catalan responses do.

These results provide strong evidence for our two hypotheses. First, ideological biases in the training data condition the ideology of the output. Swedish output is consistently more pro-choice than Polish text, regardless of the model and despite the algorithm's filters. Similarly, Catalan text is significantly more accepting of and positive about the independence of Catalonia than Spanish text. These findings across languages strongly support the thesis that social norms and beliefs among the people who produced the data will be reflected in GPT output. Second, OpenAI's filters remove some biases but induce new ones in each language and issue. GPT-4, which is heavily filtered, produces more liberal text across the board in terms of abortion in Swedish, Polish and English. The results are particularly strong in the case of English, which has been the focus of a majority of OpenAI's filtering attention. GPT-4 is almost exclusively pro-choice. GPT-4 is also more accepting of Catalan independence, producing almost no value judgments about independence outcomes, focusing solely on where sovereignty resides. Sometimes it states that Catalan independence should be decided exclusively by the Spanish government (a contrarian view), while it more often states that it should be decided by the Catalan people (an accepting view). Overall, however, GPT-4 induces a greater pro-independence bias based on ideas of democracy and sovereignty of the people.

7. Results—annotation

Table 2 displays the results of three MLM for economics, health, and both topics combined. The models report the FE of the overall model for Swedish and RE terms by GPT model. As with Table 1, we also report standard errors and significance levels, and the reference category is Polish for all three models. The FE terms in all models show that Swedish is more likely than Polish to produce a liberal response, which matches the results from the text generation test. The results are significant at the 0.001 level for all models. As for the RE terms, we see that the slope for Swedish is negative and

Table 2. Summary of multilevel models predicting biased GPT responses

	(1) Economics			(2) Health			(3) All		
	FE	*RE*		*FE*	*RE*		*FE*	*RE*	
		GPT-3.5	GPT-4		GPT-3.5	GPT-4		GPT-3.5	GPT-4
Swedish	1.075**	0.287*	−0.281*	1.308**	0.280**	−0.276**	1.210**	0.294**	−0.291**
	(0.186)	(0.102)	(0.113)	(0.252)	(0.064)	(0.078)	(0.186)	(0.056)	(0.066)
Intercept	−1.424**			−1.279*			−1.341**		
	(0.306)			(0.420)			(0.367)		
Observations	640			868			1,508		
Log Lik.	−369.881			−519.268			−890.589		

Note: ** $p \leq 0.001$, * $p \leq 0.01$, $^{+}$ $p \leq 0.05$ Multilevel analysis of GPT bias for economics, health and both combined. The reference category is Polish. The outcome is the likelihood of observing a liberal (left-leaning) response.

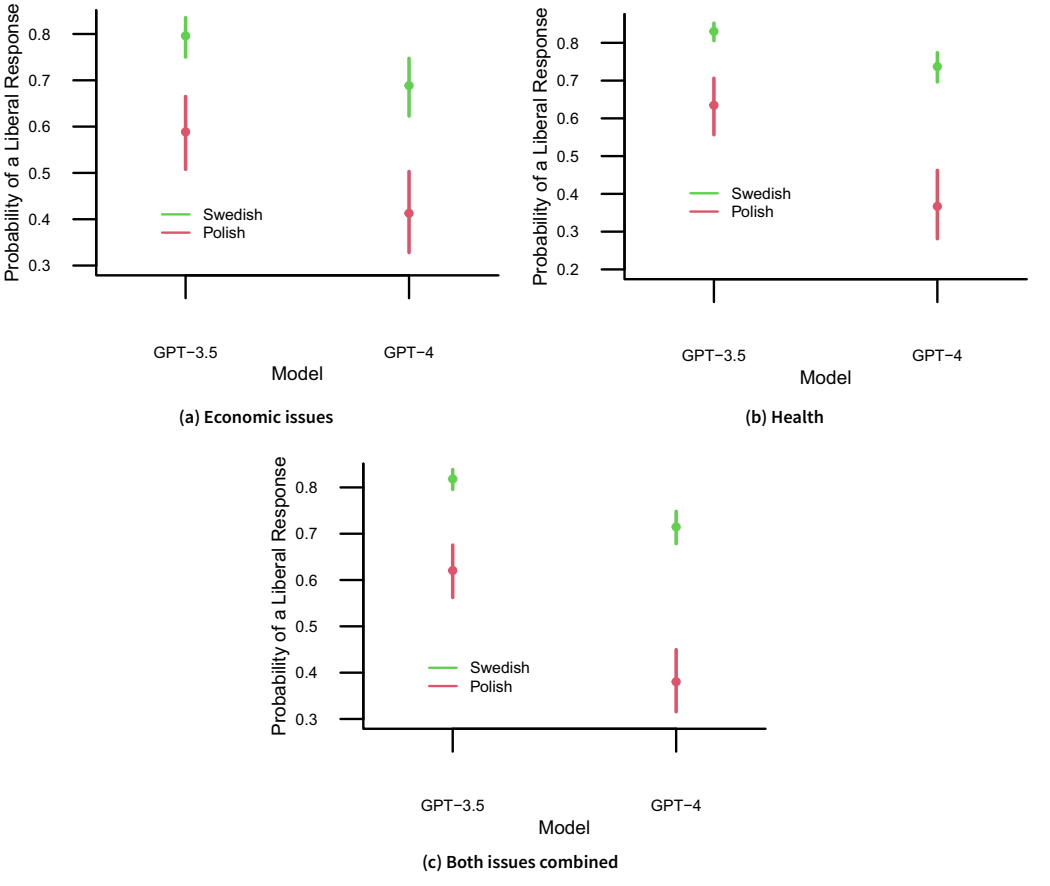

Figure 3. The predicted probabilities of observing a liberal (left-leaning) response on economic issues (a), health (b) and both (c) by GPT model and language.

statistically significant, with an overall difference of −0.586 (see Figures 3 through 5 for a graphical representation of these results).

Figure 3 confirms the results from Table 2 and shows the substantive significance of the differences across languages and models in economic issues and health policy. Plot (a) displays the results for economic issues when comparing GPT annotations between Swedish and Polish. Plot (b) plots shows the results for health while plot (c) shows the results for the combined data with both economic

issues and health policy. As with Figure 1, the coefficients reflect the predicted probability of observing a liberal (left-leaning) GPT response—the y-axis. There are two key takeaways from these results. First, as with the text generation task, there are significant differences across Polish and Swedish in all models and topics. The probability of observing a liberal response by GPT (3.5 and 4) is consistently higher in Swedish than in Polish. In economic issues (plot (a)), the probability of a liberal text is 0.588 in Polish and 0.796 in Swedish with GPT-3.5, a difference of 20.8 percentage points or 35.3%. For GPT-4, the difference is 27.6 points and 66.8% (0.689 for Swedish and 0.413 for Polish). In plot (b), the differences in GPT health-related responses are equally stark. GPT-3.5 responses are 30.7% more likely to be liberal in Swedish than in Polish,[12] while GPT-4 output is *twice* as likely to be liberal in Swedish than in Polish.[13] Lastly, the results in plot (c) where data for both issues is combined are consistent with the first two plots.[14] Therefore, the results with our language-based design show that ideological bias is significant across different LLM tasks such as text generation and annotation. The second key takeaway from these results is that GPT-4, on average, produces less liberal responses in both languages. Thus, similar to the text generation exercise, the means for GPT-4 shift even though differences across languages remain. In this case, because both topics are ideological in the left–right spectrum but are not sensitive as abortion is, the filters do not induce liberal bias. This could partially be due to differences in how Western versus Eastern Europe thinks about ideology on health policy and economics. For example, while Poland is considered more 'conservative' economically by the West for not following neoliberal ideals, in some instances it may be seen as more left-leaning following its historical ties to communist state-ownership of the means of production. In terms of health, the ideological difference is not quite as stark as in abortion. Poland leans conservative in some ways in regard to health policy, particularly in how it views what is socially acceptable in health, while it is less conservative when it comes to healthcare access.

8. Discussion and conclusion

We introduce a novel method to identify bias in generative AI models such as GPT-3.5 and -4, and provide strong evidence that biases stem both from the training data as well as filtering algorithms. Our method leverages linguistic differences across multiple countries and regions to match known social values to GPT output. Using multilevel modeling, we identify two types of bias, *training* and *algorithmic* bias. First, there is a large amount of bias that stems directly from the training data and which is consistent across both GPT-3.5 and -4. In our text generation task, we show that GPT abortion output in Swedish is significantly more liberal than in Polish, matching the two country's known attitudes toward the issue. Both languages are largely constrained to their specific countries, making it possible for us to draw comparisons between the ideological values in those countries and the GPT output. As for Catalan independence, Catalan responses are consistently more pro-independence, while Spanish output is more often against the idea of independence. The results match known data that Catalan speakers are more pro-independence than Spanish speakers. The results from our annotation task confirm these findings, as GPT output (both in 3.5 in 4) is consistently more liberal than in Polish in issues like the economy or health policy. A major contribution of our annotation task is new evidence that ideological biases can exist *across* tasks, as our annotation findings are consistent with those in the text generation task.

Second, we find that OpenAI's filtering induces liberal, pro-choice biases in GPT-4 responses in our text generation task with two politically sensitive topics. Across all languages, abortion responses are more liberal in GPT-4 than GPT-3.5. For Polish and Swedish (see Figure 1), GPT-4 responses

[12]The probability is 0.830 for Swedish and 0.635 for Polish.

[13]The probability is 0.737 for Swedish and 0.367 for Polish, a 100.8% increase.

[14]In plot (c), the probability of a liberal response in Swedish with GPT-3.5 is 0.818. It is 0.621 in Polish. The difference is 19.7 percentage points and 31.7%. For GPT-4, the respective probabilities are 0.715 in Swedish and 0.380 in Polish, for a difference of 33.5 percentage points and 88.2%.

are 30.4% and 25.5% more liberal, respectively. For English, they are 94% more liberal, and GPT-4 produces liberal text 95.9% of the time. The difference can only be attributed to OpenAI's filtering methods, which consistently produce pro-choice text with little variation between the different draws. A similar pattern emerges with Catalan independence. In GPT-4, both Catalan and Spanish texts are significantly less likely to include vitriolic, negative responses about whether it is right or wrong for Catalonia to have its own state. Neither state that independence would be 'illegal,' 'a catastrophe,' or 'an abject fiasco.' Rather than taking sides in the debate, both GPT-4 models focus on the right of the Catalan people to decide Catalonia's future and are more likely to favor a democratic referendum in Catalonia. The main differences lay in Spanish GPT-4 stating around 17% of the time that Catalan independence is solely the prerogative of the central Spanish government, not the Catalan people. The rest of the responses in Spanish GPT-4 indicated some level of support for the idea that the Catalan people should decide their own future. Therefore, both GPT-4 models are much more liberal and pro-choice. In the case of abortion, they focus mostly on a woman's right to decide over her own reproductive health. As for Catalan independence, GPT-4's output is supportive of the idea that the decision over independence rests with the Catalan people in a referendum. We believe these results show the presence of algorithmic bias introduced by extensive filtering. Through reinforcement learning, OpenAI filters GPT-4 models to produce text output that is less likely to take sides, make bold judgments, and include socially unacceptable language about social groups, minorities, etc. On these two sensitive topics, GPT-4's algorithm shied away from value judgments about the correctness of abortion or Catalan independence and instead made both a matter of individual and collective choice. GPT-3.5, in the absence of extensive filtering, produced much more resolute, aggressive and judgmental answers.

The contributions of this work are **many**. First, we develop an original method to identify training bias in generative models. Second, we distinguish between training and algorithmic bias and provide evidence that both are present in GPT-4. Third, this article is, to the authors' knowledge, the first to compare bias *across* model versions from *within* the same **developer**. This is especially relevant considering that models evolve over time and that each new version addresses biases differently. Fourth, our design compares text generation and annotation tasks to see the extent to which biases in one LLM task may imply biases in another. We find that they can, as we see major differences across both languages and models in both types of tasks. Lastly, our work has major implications for the politics of AI. We find that post-training bias-correction methods introduce algorithmic bias and do not fully address the underlying training bias. Most concerning is that these approaches, in fact, introduce new biases. Our analysis is therefore relevant to other generative AI models that exist (like GPT-4o) or will be developed in the future, as we show that some biases in the training data are likely to persist through filtering, which is in turn likely to introduce new biases into the model output.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2025.10057. To obtain replication material for this article, https://doi.org/10.7910/DVN/NYRTCA.

Acknowledgements. The authors thank Kaylyn Schiff, Sebastián Vallejo Vera and Bryce Dietrich for their helpful comments and suggestions on previous versions of the paper. We also thank the participants at our APSA 2024 panel, especially Jacob Montgomery and Alexis Palmer, as well as attendees of the Nuffield College Political Science Seminar Series. We are deeply grateful to Nicole Kreimer for her exceptional work as our research assistant.

Data and code availability statement. The labeled data generated from GPT-3.5 and -4 and the replication code are available at https://github.com/joantimoneda/PSRM_GPT_bias

Competing interests. None.

References

Abid A, Farooqi M and Zou J (2021) Large language models associate Muslims with violence. *Nature Machine Intelligence* **3**, 461–463.

Andersson PF (2022) Taxation and left-wing redistribution: The politics of consumption tax in britain and sweden. *Comparative Politics* **54**, 279–301.

Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C and Wingate D (2023) Out of one, many: Using language models to simulate human samples. *Political Analysis* **31**, 337–351.

Atienza-Barthelemy J, Martin-Gutierrez S, Losada JC and Benito RM (2019) Relationship between ideology and language in the Catalan independence context. *Scientific Reports* **9**, 1–13.

Bluhm K and Varga M (2020) Conservative developmental statism in east central europe and russia. *New Political Economy* **25**, 642–659.

Buchholz, MG (2023) Assessing the effectiveness of gpt-3 in detecting false political statements: A case study on the liar dataset. *arXiv preprint*.

Carvalho A (2007) Ideological cultures and media discourses on scientific knowledge: Re-reading news on climate change. *Public Understanding of Science* **16**, 223–243.

Chang CT, Srivathsa N, Bou-Khalil C, Swaminathan A, Lunn MR, Mishra K, Koyejo S and Daneshjou R (2025) Evaluating anti-lgbtqia+ medical bias in large language models. *PLOS Digital Health* **4**, p.e0001001.

Collins PA, Abelson J and Eyles JD (2007) Knowledge into action?: Understanding ideological barriers to addressing health inequalities at the local level. *Health Policy* **80**, 158–171.

Cooper K (2023) *OpenAI GPT-3: Everything You Need to Know*, Section: Data Science. *Springboard Blog*.

De Vries CE and Giger N (2014) Holding governments accountable? Individual heterogeneity in performance voting. *European Journal of Political Research* **53**, 345–362.

Doering J (2014) A Battleground of Identity: Racial Formation and the African American Discourse on Interracial Marriage. *Social Problems* **61**, 559–575.

Fetterlorf J and Clancy L (2024) Support for legal abortion is widespread in many countries, especially in europe. *Pew Research*.

Gelman A (2006) Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* **48**, 432–435.

Gilardi F, Alizadeh M and Kubli M (2023) Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* **120**, e2305016120.

Haini H and Loon PW (2021) Does government ideology affect the relationship between government spending and economic growth?. *Economic Papers: A Journal of Applied Economics and Policy* **40**, 209–216.

Hazlett C and Wainstein L (2022) Understanding, choosing, and unifying multilevel and fixed effect approaches. *Political Analysis* **30**, 46–65.

Heikkilä M (2023) How OpenAI is trying to make ChatGPT safer and less biased. *MIT Technology Review*.

Hernández E and Kriesi H (2016) The electoral consequences of the financial and economic crisis in europe. *European Journal of Political Research* **55**, 203–224.

Holsti OR. 1969. *Content analysis for the social sciences and humanities. Reading, Addison-Wesley (content analysis), MA*.

Ibrahim, H., F. Khan, H. Alabdouli, M. Almatrooshi, T. Nguyen, T. Rahwan, Y. Zaki (2024) Analyzing political stances on twitter in the lead-up to the 2024 us election. *arXiv preprint arXiv:2412.02712*.

Jin, X., F. Barbieri, B. Kennedy, A. M. Davani, L. Neves, X. Ren (2020) On transferability of bias mitigation effects in language model fine-tuning. *arXiv preprint arXiv:2010.12864*.

Kelly W (2024) GPT-3.5 vs. GPT-4: Biggest differences to consider. *Enterprise AI*.

Klann EM and Wong YJ (2020) A Pregnancy Decision-Making Model: Psychological, Relational, and Cultural Factors Affecting Unintended Pregnancy. *Psychology of Women Quarterly* **44**, 170–186.

Koralewska I and Zielińska K (2022) 'Defending the unborn', 'protecting women' and 'preserving culture and nation': Anti-abortion discourse in the Polish right-wing press. *Culture, Health & Sexuality* **24**, 673–687.

Kotek H, Dockum R and Sun D (2023) Gender bias and stereotypes in Large Language Models. In Proceedings of The ACM Collective Intelligence Conference, CI'23. Association for Computing Machinery, pp. 12–24. New York, NY, USA, pp.

Le Mens G, Kovács B, Hannan MT and Pros G (2023) Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences* **120**, e2309350120.

Lee K, Paci S, Park J, You HY and Zheng S (2024) Applications of gpt in political science research. *PS: Political Science and Politics*.

Lee S, Peng T-Q, Goldberg MH, Rosenthal SA, Kotcher JE, Maibach EW and Leiserowitz A (2024) Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLOS Climate* **3**, e0000429.

Llaneras K (2017) Income and origins sway support for independence. *El País*.

Lucy L and Bamman D (2021) Gender and Representation Bias in GPT-3 Generated Stories. In Proceedings of the Third Workshop on Narrative Understanding, Virtual. Association for Computational Linguistics, pp. 48–55.

Lunardi R, Barbera DL and Roitero K (2024) The elusiveness of detecting political bias in language models. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 3922–3926.
Lupo L, Magnusson O, Hovy D, Naurin E and Wängnerud L (2023) How to use large language models for text coding: The case of fatherhood roles in public policy documents. arXiv preprint arXiv:2311.11844.
Mellon J, Bailey J, Scott R, Breckwoldt J, Miori M and Schmedeman P (2024) Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics* **11**, 1–7.
Moon DS, Thompson J and Whiting S (2019) Lost in the Process? The impact of devolution on abortion law in the United Kingdom. *The British Journal of Politics and International Relations* **21**, 728–745.
Motoki F, Neto VP and Rodrigues V (2024) More human than human: Measuring ChatGPT political bias. *Public Choice* **198**, 3–23.
O'Hagan S and Schein A (2024) Measurement in the age of LLMs: An Application to ideological scaling. arXiv preprint arXiv:2312.09203.
OpenAI (2024). GPT-4 Technical Report.
Ornstein JT, Blasingame EN and Truscott JS (n.d) How to train your stochastic parrot: Large language models for political texts.
Peterson MA (2011) The ideological and partisan polarization of healthcare reform and tax policy. *Tax Law Review* **65**, 627.
Pienkos D (2024) Poland, American Polonia, and Poland's Borders: Another Way to Understand the Connnection?. *Kuryer Polski*.
Pit P, Ma X, Conway M, Chen Q, Bailey J, Pit H, Keo P, Diep W and Jiang Y-G (2024) Whose side are you on? Investigating the political stance of large language models. arXiv preprint arXiv:2403.13840.
Ray PP (2023) Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* **3**, 121–154.
Rodriguez C and Ditto PH (2020) Beyond Misogyny: Sexual Morality and Sanctity of Life predict Abortion Attitudes.
Roemer G, Li A, Mahmood U, Dauer L and Bellamy M (2024) Artificial intelligence model GPT4 narrowly fails simulated radiological protection exam. *Journal of Radiological Protection* **44**, 013502.
Santurkar S, Durmus E, Ladhak F, Lee C, Liang P and Hashimoto T (2023) Whose opinions do language models reflect?. In International Conference on Machine Learning (pp. 29971–30004). PMLR.
Schiff DS, Schiff KJ and Pierson P (2022) Assessing public value failure in government adoption of artificial intelligence. *Public Administration* **100**, 653–673.
Sheng E, Chang K-W, Natarajan P and Peng N (2019) The woman worked as a babysitter: On biases in language generation. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 3407–3412).
Si C, Gan Z, Yang Z, Wang S, Wang J, Boyd-Graber J and Wang L (2022) Prompting gpt-3 to be reliable. *arXiv preprint*.
Srivastava S (2023) Algorithmic governance and the international politics of big tech. *Perspectives on Politics* **21**, 989–1000.
Stegmueller D (2013) How many countries for multilevel modeling? A comparison of frequentist and bayesian approaches. *American Journal of Political Science* **57**, 748–761.
Swigart KL, Anantharaman A, Williamson JA and Grandey AA (2020) Working while liberal/conservative: A review of political ideology in organizations. *Journal of Management* **46**, 1063–1091.
Sydsjö A, Josefsson A, Bladh M and Sydsjö G (2011) Trends in induced abortion among Nordic women aged 40-44 years. *Reproductive Health* **8**, 23.
Timoneda JC and Vallejo Vera S (2025a) BERT, RoBERTa, or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text. *The Journal of Politics* **87**, 347–364.
Timoneda JC and Vallejo Vera S (2025b) Behind the mask: Random and selective masking in transformer models applied to specialized social science texts. *PloS one* **20**, e0318421.
Vallgårda S (2007) Public health policies: A scandinavian model?. *Scandinavian Journal of Public Health* **35**, 205–211.
Wu PY, Nagler J, Tucker JA and Messing S (2023) Large language models can be used to estimate the latent positions of politicians. arXiv preprint arXiv:2303.12057.
Young FI, Sullivan D and Hamann HA (2020) Abortions due to the Zika virus versus fetal alcohol syndrome: Attributions and willingness to help. *Stigma and Health* **5**, 304–314.
Zabdyr-Jamróz M, Löblová O, Moise AD and Kowalska-Bobko I (2021) Is the polish 'law and justice'(pis) a typical populist radical right party? a health policy perspective. *The Populist Radical Right and Health: National Policies and Global Trends* 113–137.
Zmigrod L (2020) A psychology of ideology: Unpacking the psychological structure of ideological thinking. *Perspectives on Psychological Science* **17**, 1072–1092.

Cite this article: Walker CP and Timoneda JC (2025) Is ChatGPT conservative or liberal? A novel approach to assess ideological stances and biases in generative LLMs. *Political Science Research and Methods*, 1–15. https://doi.org/10.1017/psrm.2025.10057