Are Large Language Models Chameleons?

Mingmeng Geng [1] **Sihong He** [2] **Roberto Trotta** [1 3]

Abstract

Do large language models (LLMs) have their own worldviews and personality tendencies? Simulations in which an LLM was asked to answer subjective questions were conducted more than 1 million times. Comparison of the responses from different LLMs with data from the European Social Survey (ESS) suggests that the effect of prompts on bias and variability is fundamental, highlighting major cultural, age, and gender biases. Methods for measuring the difference between LLMs and survey data are discussed, such as calculating weighted means and a new proposed measure inspired by Jaccard similarity. We conclude that it is important to analyze the robustness and variability of prompts before using LLMs to model individual decisions or collective behavior, as their imitation abilities are approximate at best.

1. Introduction

Although often referred to as "stochastic parrots" (Bender et al., 2021), large language models (LLMs) can do more than simply "talk" (Staab et al., 2023). Their adaptability and versatility remind us of another animal: the chameleon. By analyzing whether LLMs are able to play a specific human role as required, we hope to gain inspiration for analyzing the cognitive abilities of LLMs.

Do LLMs have their own opinions? If we ask LLMs to answer some subjective questions, will they behave like humans?

The alignment between LLMs outputs and human values has attracted great interest and has already been measured (Hendrycks et al., 2020a). What could be called the "personality traits" of LLMs have also been widely explored in several papers (Safdari et al., 2023; Jiang et al., 2023b; 2024; Mei et al., 2024). Comparing the responses of LLMs to subjective questions with those provided by human beings can be achieved in multiple ways (Durmus et al., 2023; Takemoto, 2024); importantly, human opinions are not uniformly distributed among people, and indeed this is the crux of reliable surveys.

What if prompts with personal information are used to make LLMs' responses more specific?

In survey research, a number of demographic variables are routinely used, such as gender, race, and age (Goldberg et al., 1998; Herek, 2002; Kyriakidis et al., 2015). Gender and race are two important perspectives to consider for bias in LLMs (Sun et al., 2023; Kotek et al., 2023; Dong et al., 2024). The sensitivity of LLMs to age was also analyzed, based on six age groups (Liu et al., 2024). But the influence and robustness of the prompts are rarely discussed in papers dealing with opinions expressed by LLMs (Röttger et al., 2024).

How can we quantitatively analyze and compare the trustworthiness of LLMs by means of survey data?

Bias in LLMs is almost impossible to avoid, even with subsequent adjustments and alignments (Navigli et al., 2023). Our goal is to measure the difference between LLMs' responses and what real people think, by asking LLMs to answer actual survey questions with a prompt giving the corresponding demographic variables. We aim to then compare similarities and differences between the simulated results and actual survey data. The data and results obtained from surveys also deviate to a greater or lesser extent from the "ground truth", while the representativeness of the sample is often debated (Russell et al., 2022; Grafström and Schelin, 2014). Therefore, we want to analyze the bias in responses of LLMs, using real survey data as the "ground truth", which means that the survey results are assumed to be representative.

2. Data

To better compare with real people's opinions, we made use of data from the European Social Survey (ESS), a biennial survey of attitudes and behaviour. We chose the round 10 of ESS (European Social Survey ERIC (ESS ERIC), 2022),

[1]International School for Advanced Studies (SISSA), Italy [2]University of Texas at Arlington, USA [3]Imperial College London, UK. Correspondence to: Mingmeng Geng <mgeng@sissa.it>, Roberto Trotta <rtrotta@sissa.it>.

The 41st International Conference on Machine Learning (ICML) Workshop on LLMs and Cognition, Vienna, Austria.

collected from September 2020 up until May 2022.

3. Methods

3.1. Simulations

The performance of LLMs needs to be evaluated from simulations, as it is not possible to predict the response of LLMs through analysis of the model structure and its parameters. We investigated ChatGPT (including *GPT-3.5* and *GPT-4o*), as well as open source models via APIs or locally, including LLaMA-2 (Touvron et al., 2023), LLaMA-3, Mistral (Jiang et al., 2023a) and DeepSeek-V2 (Bi et al., 2024). As with the real survey, we used the *Chat* model rather than the completion of sentences. For example:

P1 *You are a [man or woman] born in [year] living in [country]. To what extent do you agree with the following statement: '[statement]'? Please respond with only a number (1) Agree strongly (2) Agree (3) Neither agree nor disagree (4) Disagree (5) Disagree strongly.*

In real surveys, the order in which questions are asked affects the respondents' answers (Carlsson et al., 2012). We ask one question at a time, so each answer is zero-shot. The same personal information as in the real survey (nationality, gender, and age) was used in the simulations.

3.2. Measurements

Previous studies have shown that LLMs are likely to be biased, but there is no agreement on how large these biases are. Theoretically, this depends not only on the model itself but also on the questions being asked.

We are dealing with subjective questions that do not have standard answers, but the results of surveys from real people are available. Therefore, we focus on data where participants agree or disagree with a statement, while for simplicity the missing data options ("Refusal", "Don't know", "No answer") were disregarded. The ordinal numbers of the responses from 1 ("Agree strongly") to 5 ("Disagree strongly") were used as a scale for their level of disagreement.

In orde to compare responses from LLMs to the ones from humans, we want to evaluate not only averages, but also their variance. To better compare the impact of different prompts and different LLMs, we employ the following measures: mean, variance, and J-index.

Inspired by Jaccard similarity (Chierichetti et al., 2010), we define the J-index $J_q(G)$ for statement q and people group G that we are interested in (e.g., a country) as:

$$J_q(G) = \frac{\sum_{g\in G} I\left(R_q^h(g), R_q^m(g)\right)}{\sum_{g\in G} U\left(R_q^h(g), R_q^m(g)\right)} \quad (1)$$

where $I(\cdot,\cdot)$ and $U(\cdot,\cdot)$ are the intersection and union of two sets of responses, g is the subgroup of G, $R_q^h(g)$ and $R_q^m(g)$ represent the responses from survey data and LLM simulations, respectively. A J-index of 1 thus corresponds to perfect congruence between between survey and LLM simulation.

We further take the analysis weight (combination of post-stratification weight and population weight) w_i for participant i provided in the ESS dataset to re-weigh $I\left(R_q^h(g), R_q^m(g)\right)$ and $U\left(R_q^h(g), R_q^m(g)\right)$ to improve the representativeness of survey respondents concerning the target population,

$$\begin{aligned} & I\left(R_q^h(g), R_q^m(g)\right) \\ &= \sum_{v\in V_q}\left(\min\left(\sum_{i\in g} w_i \mathbb{1}_{r_q^h(i)=v}, \sum_{i\in g} w_i \mathbb{1}_{r_q^m(i)=v}\right)\right) \end{aligned} \quad (2)$$

$$\begin{aligned} & U\left(R_q^h(g), R_q^m(g)\right) \\ &= \sum_{v\in V_q}\left(\max\left(\sum_{i\in g} w_i \mathbb{1}_{r_q^h(i)=v}, \sum_{i\in g} w_i \mathbb{1}_{r_q^m(i)=v}\right)\right) \end{aligned} \quad (3)$$

where v corresponds to the value of the response, V_q is the set of all possible values for question q, $r_q^h(i)$ represents the answer in survey data, and $r_q^m(i)$ means the response of LLM given the same information as participant i.

4. Results

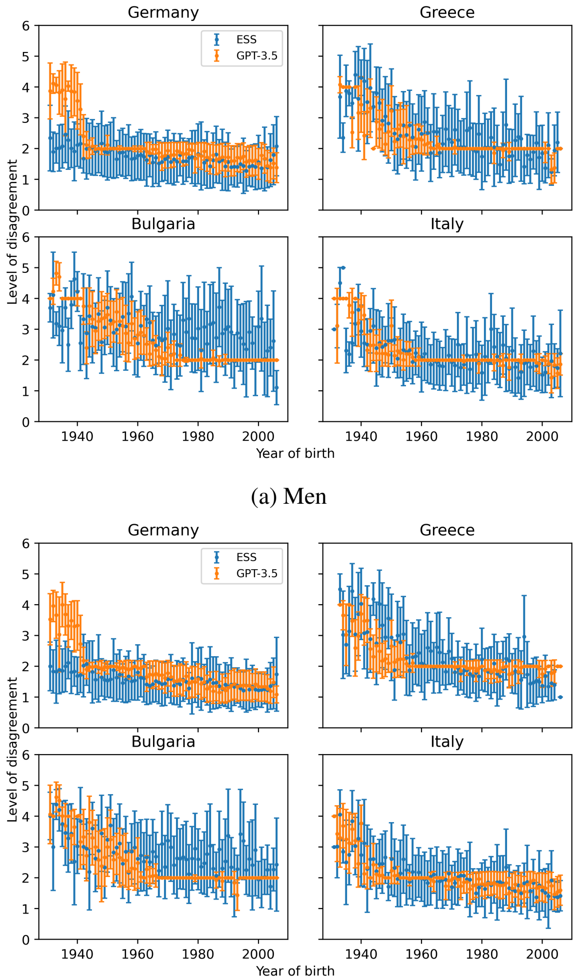
In our analysis, we used the four countries with the highest number of participants in the ESS (Germany, Greece, Bulgaria, and Italy, with 16132 samples in total). All of our results take post-stratification weighting into account.

4.1. ChatGPT simulations

Considering answers from *GPT-3.5* [1] and using the prompt **P1** as described above, we adopted specific birth years rather than age groupings to better show the sensitivity of ChatGPT to age information. ChatGPT does not always respond with only one number as we requested, sometimes including a description after the number (e.g., "(1) Agree strongly"), but it doesn't refuse to answer.

The results for the first statement are shown in Figure 1. At least on this question, the output of *GPT-3.5* is indeed determined by the demographic variables given in the prompts. The mean of the simulation results is close to the mean of ESS data, except for Germans born before 1940. However, the variance of the simulated data is too small compared to the real data, even when the *temperature* parameter is set to its maximum value.

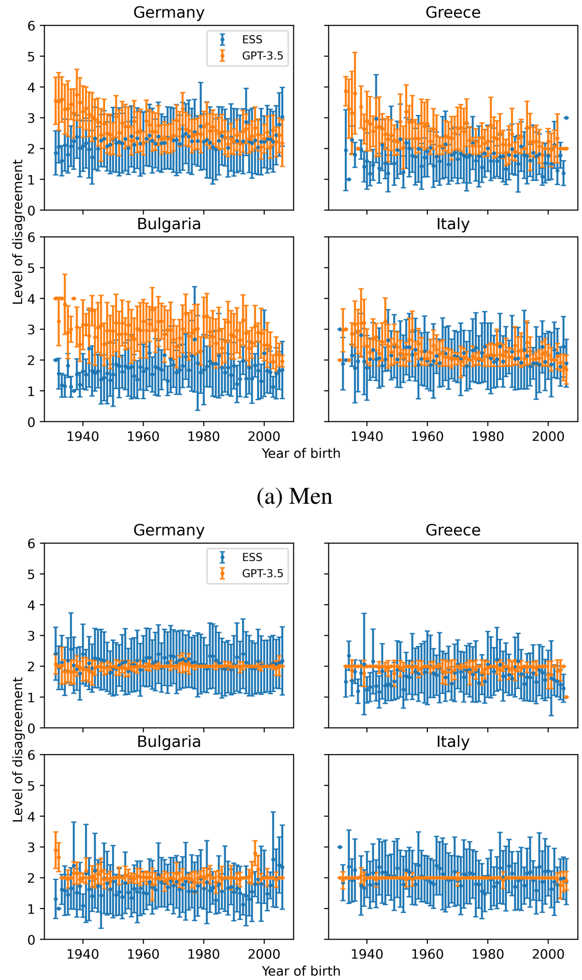
[1] gpt-3.5-turbo-0125, temperature = 1, top_p = 0.9



(a) Men

(b) Women

Figure 1: **Q1**: "Gays and lesbians free to live life as they wish"? Prompt: **P1**. The points represent the mean and the error bars represent the standard deviation (and the same for the next figures). Model: *GPT-3.5*.



(a) Men

(b) Women

Figure 2: **Q2**: "Government should reduce differences in income levels"? Prompt: **P1**. Model: *GPT-3.5*.

The impact of the same variable can vary for different topics, e.g, the difference resulting from gender is evident in LLMs' response to the second question, as seen in Figure 2.

Although we always used different random seeds for different persons and the same random seed for the same person across different questions, the problem of too little variance in the simulations is common for various questions, as has been pointed out by others (Boelaert et al., 2024).

In particular, the example in Figure 2 shows that although the mean of ESS data and *GPT-3.5* simulations are close for women, the variance of simulations is very small –almost non-existent– in this case. This would seem to imply that women are associated with more stereotypical answers, another facet of gender bias.

4.2. Prompts and parameters

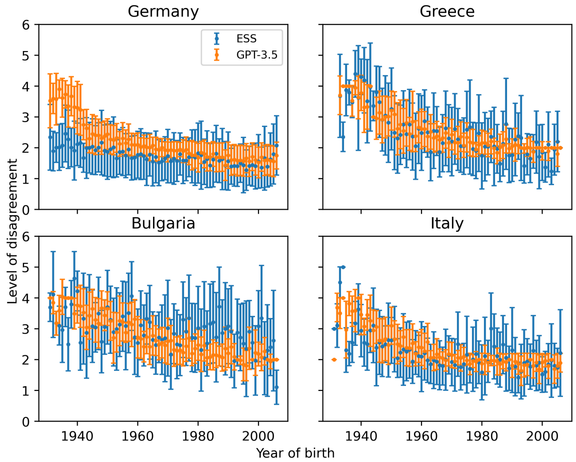
In addition to the biases and stereotypes that may result from the model itself and the training data, the issue related to variance is also likely caused by the fact that we provided too little information in prompt **P1**. Therefore, we tried prompts with a more informative description, such as the respondent's occupation:

P2 *You are a [man or woman] born in [year] living in [country]. **Occupation category: [ISCO[2]]**...*

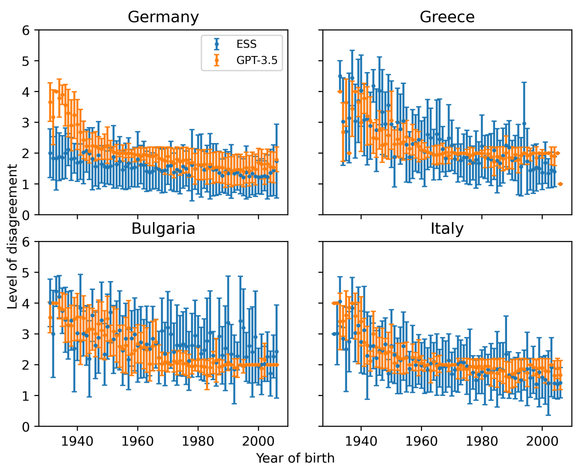
The results generated by prompt **P2** are shown in Figure 3, where the variability of simulations becomes closer to the variance of survey data. This is not because our addition to the prompt has increased the degree of random-

[2] International Standard Classification of Occupations (ISCO)

ness, but rather because occupational differences were previously ignored, suggesting that a "regression to the mean" phenomenon was at play with prompt **P1**. The impact of prompts goes beyond how much information is provided – the order also matters, as has been shown previously (Pezeshkpour and Hruschka, 2023). For example, we tried reversing the order of answer options:



(a) Men



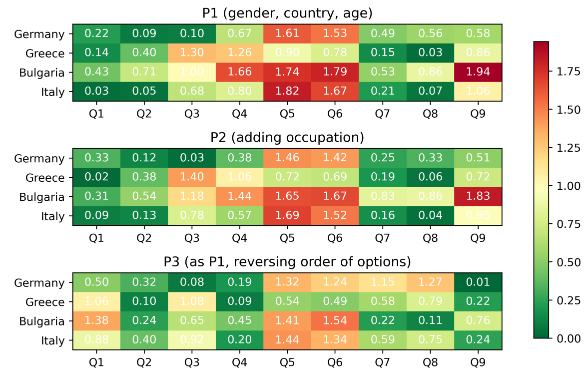
(b) Women

Figure 3: **Q1**: "Gays and lesbians free to live life as they wish"? Prompt: **P2**. Model: *GPT-3.5*.

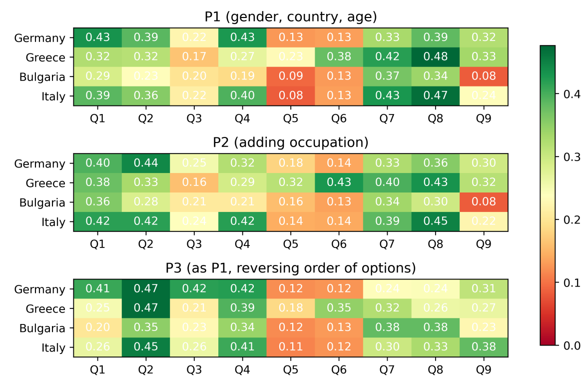
P3 . . . ***(1) Disagree strongly (2) Disagree (3) Neither agree nor disagree (4) Agree (5) Agree strongly.***

It is well known that the order in which options are present affects humans' choice (Galesic et al., 2008). The same appears to be true for LLMs, to an even larger extent at times, such as the simulation results using **P3** for **Q1**, as shown in Figure 6 of the appendix.

For 9 such questions in the ESS (listed in the appendix), we calculated the mean bias and J-index for the simulations with *GPT-3.5* compared to ESS data, and the results are shown in Figure 4. For ease of plotting, we show the absolute value of the mean difference in the figure.



(a) Absolute value of the mean difference (error).



(b) J-index (congruence).

Figure 4: Comparisons between survey data and simulation results based on *GPT-3.5*.

As we have seen above, prompt **P3** leads to markedly different simulation results for **Q1**, although this does not always hold true for other questions. For example, the simulation results for **Q4** using prompt **P3** produced better results than those using prompts **P1** and **P2**. The J-index brings further insight. For example, while **Q3** is still the worst question to simulate for Greeks, it does not simulate well for the Germans, either.

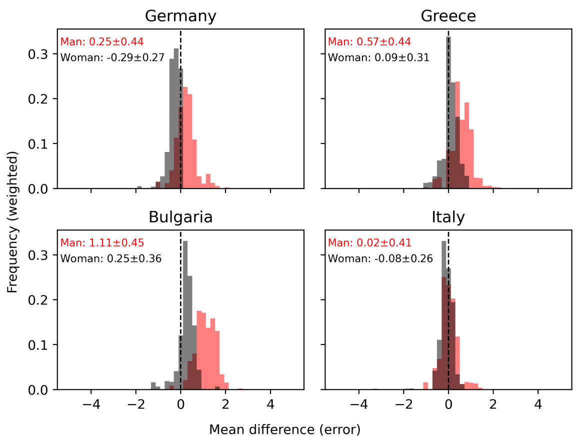
4.3. Other LLMs

Unlike ChatGPT, the other LLMs may refuse to answer, and their responses are not always valid (see appendix for examples), which could be considered as missing data. We could force them to answer by changing prompts, but this would also affect their choices (Röttger et al., 2024). Thus we use the same prompts **P1** and **P2** as before. For a more quantiative comparison, we calculated the difference of the mean and J-index compared to ESS data for **Q1**, listed in

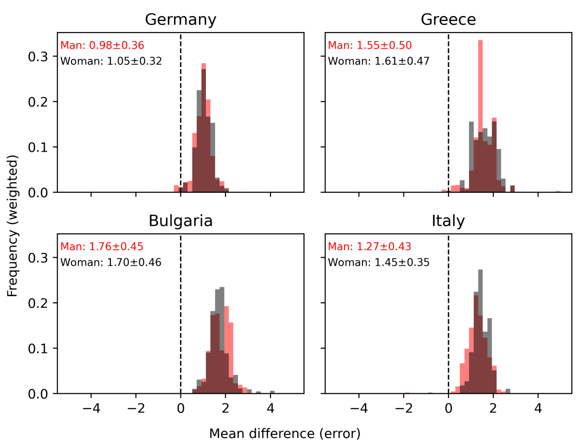
Table 1 and Table 2 in the appendix for 7 different LLMs.

For instance, Table 1 shows that *DeepSeek-V2* and *GPT-3.5* generate simulation results with weighted averages closest to the survey data, while *LLaMA-3-8B* is highly biased for both prompts; occupation information is very important for *GPT-4o*, which appeats to perform generally worse than its predecessor *GPT-3.5*, which is somewhat unexpected.

People with the same country, gender, and year of birth are considered to be in the same group. Figure 5 compares the mean difference of each gender.



(a) Model: *GPT-3.5*



(b) Model: *LLaMA-2-7B*

Figure 5: **Q2**: "Government should reduce differences in income levels"? Prompt: **P1**.

While *LLaMA-2-7B* shows a larger bias, it appears to be almost gender-independent. *GPT-3.5* shows an overall smaller bias, but a more marked influence of gender in both mean and standard deviation.

5. Conclusions

Can a LLM adjust itself to impersonate a human like a chameleon? It depends on the object of imitation and the capabilities of the given LLM. Regardless of the goals and process of LLM alignment, our results reflect a clear geographic imbalance.

We proposed and validated possible ways to make LLM responses more human-like, such as providing additional information in prompts. We also presented a metric for LLM responses to subjective questions that focuses more on the distribution than the mean, and crucially, suggested to compare with actual survey data in order to highlight potential gender and cultural biases in the response. The same techniques can be used in the analysis of cognitive abilities.

Although LLMs show the potential to perform simulations of collective behavior or to replace human participants in limited settings, more advanced LLMs do not necessarily produce simulation results that are more similar to survey data. On the contrary, earlier vision LLMs may give better simulations on some subjective questions.

6. Discussions

We know that LLMs can infer people's gender, age, and location from redacted and anonymized input text (Staab et al., 2023). Conversely, prompting LLMs with appropriate demographic variables as we have done here produces answers that appear to be generally aligned with real people's views, although still suffering from biases and stereotypes.

We used survey data for comparison, which should be more representative of public opinion than crowd-sourced data. Compared to previous papers, more detailed simulations and analysis are performed, such as considering precise age information and using the weights of the survey data to better estimate the target population. Our approach could be further improved, e.g. in the handling of missing data.

LLMs have demonstrated the ability and potential to transform the field of computational social science (Ziems et al., 2024). A completely neutral LLM is not necessarily a good aim. A model with higher ratings or more parameters is not necessarily more suitable for survey simulation for a number of reasons, such as subsequent human adjustments. And the fine-tuning steps may well remain absolutely necessary.

The same choices, challenges, and limits may be present when dealing with cognitive tasks with LLMs. We need to analyze carefully, for example, the robustness of prompts before simulating human behavior with LLMs.

Limitations

Due to budgetary and computational resource constraints, we "only" performed about 1 million simulations, although more simulations would be beneficial in obtaining more complete conclusions, for example for investigating how the language used in the prompt may affect the outcome (AlKhamissi et al., 2024). LLMs are updated exceptionally quickly, and comparing which model is better can easily become outdated, so it's important to explore commonalities in terms of opinions in these LLMs. In addition, we are currently considering only 4 countries and 9 questions, which could be expanded to obtain a fuller picture of how European cultural diversity is captured (if at all) in the current generation of LLMs.

Acknowledgments

The authors would like to thank the personnel of RACHAEL S.r.l. for helpful discussions. R.T. acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 – Project FAIR "Future Artificial Intelligence Research". M.G. is supported by DM Dottorati Innovazione e Green ex DM1061 Anno 2021 (DM 1061 del 10/08/2021).

References

AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., and Diab, M. (2024). Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al. (2024). Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Boelaert, J., Coavoux, S., Ollion, É., Petev, I., and Präg, P. (2024). Machine bias generative large language models have a worldview of their own.

Carlsson, F., Mørkbak, M. R., and Olsen, S. B. (2012). The first time is the hardest: A test of ordering effects in choice experiments. *Journal of Choice Modelling*, 5(2):19–37.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., et al. (2024). Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Chierichetti, F., Kumar, R., Pandey, S., and Vassilvitskii, S. (2010). Finding the jaccard median. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 293–311. SIAM.

Dong, X., Wang, Y., Yu, P. S., and Caverlee, J. (2024). Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*.

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. (2023). Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

European Social Survey ERIC (ESS ERIC) (2022). European social survey (ESS), round 10 - 2020.

Galesic, M., Tourangeau, R., Couper, M. P., and Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public opinion quarterly*, 72(5):892–913.

Goldberg, L. R., Sweeney, D., Merenda, P. F., and Hughes Jr, J. E. (1998). Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptions of personality attributes. *Personality and Individual differences*, 24(3):393–403.

Grafström, A. and Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41(2):277–290.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. (2020a). Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020b). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Herek, G. M. (2002). Gender gaps in public opinion about lesbians and gay men. *Public opinion quarterly*, 66(1):40–66.

Hovy, D. and Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023a). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., and Zhu, Y. (2024). Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Jiang, H., Zhang, X., Cao, X., Kabbara, J., and Roy, D. (2023b). Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. (2023). Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*.

Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., and Kang, D. (2023). Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., Barhoum, A., Nguyen, D., Stanley, O., Nagyfi, R., et al. (2024). Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Kotek, H., Dockum, R., and Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Kyriakidis, M., Happee, R., and De Winter, J. C. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology and behaviour*, 32:127–140.

Liu, S., Maturi, T., Shen, S., and Mihalcea, R. (2024). The generation gap: Exploring age bias in large language models. *arXiv preprint arXiv:2404.08760*.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Cheng, R. G. H., Klochkov, Y., Taufiq, M. F., and Li, H. (2023). Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121.

Motoki, F., Pinho Neto, V., and Rodrigues, V. (2024). More human than human: Measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Navigli, R., Conia, S., and Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pezeshkpour, P. and Hruschka, E. (2023). Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., and Hovy, D. (2024). Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Russell, A. M., Browne, M., Hing, N., Rockloff, M., and Newall, P. (2022). Are any samples representative or unbiased? reply to pickering and blaszczynski. *International Gambling Studies*, 22(1):102–113.

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., and Matarić, M. (2023). Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

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*, pages 29971–30004. PMLR.

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., and Xiong, D. (2023). Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Staab, R., Vero, M., Balunović, M., and Vechev, M. (2023). Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Sun, H., Pei, J., Choi, M., and Jurgens, D. (2023). Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.

Takemoto, K. (2024). The moral machine experiment on large language models. *Royal Society open science*, 11(2):231393.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wolf, Y., Wies, N., Avnery, O., Levine, Y., and Shashua, A. (2023). Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Ye, F., Yang, M., Pang, J., Wang, L., Wong, D. F., Yilmaz, E., Shi, S., and Tu, Z. (2024). Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

A. European Social Survey (ESS) data

A.1. Statements (questions)

1. *Gays and lesbians free to live life as they wish.*

2. *Government should reduce differences in income levels.*

3. *Gay and lesbian couples right to adopt children.*

4. *Ashamed if close family member gay or lesbian.*

5. *Obedience and respect for authority most important virtues children should learn.*

6. *Country needs most loyalty towards its leaders.*

7. *Small secret group of people responsible for making all major decisions in world politics.*

8. *Groups of scientists manipulate, fabricate, or suppress evidence in order to deceive the public.*

9. *COVID-19 is result of deliberate and concealed efforts of some government or organisation.*

A.2. Data release time

1. June 2022: Bulgaria, Croatia, Czechia, Estonia, Finland, France, Hungary, Lithuania, Slovakia and Slovenia [3].

2. November 2022: Greece, Iceland, Italy, Montenegro, Netherlands, North Macedonia, Norway, Portugal, Switzerland, Austria, Germany, Poland, Serbia, Spain and Sweden[4].

3. May 2023: Belgium, Ireland, Israel, Latvia and the United Kingdom[5].

B. Related work

LLM bias Bias in natural language processing (NLP) arises for several reasons, such as data, models, and research design (Hovy and Prabhumoye, 2021). The risks and shortcomings of LLMs were addressed even before they became popular (Bender et al., 2021). While LLMs solve many of the difficulties that traditional NLP methods have failed to overcome, the challenge of bias persists (Navigli et al., 2023). For example, results based on GPT-3 show that the opinions of certain demographic groups in the United States are not well represented (Santurkar et al., 2023). The political bias of ChatGPT in other countries has also been quantified (Motoki et al., 2024). Bias also looks like a problem shared by other LLMs (Boelaert et al., 2024), and not specific to ChatGPT.

LLM evaluation The pervasiveness and influence of LLMs is increasing rapidly, and it is becoming more important to better evaluate and compare different LLMs (Chang et al., 2024). Platforms that build on users' comparative ratings have also emerged, for example, *Chatbot Arena* (Chiang et al., 2024). Traditional objective evaluation criteria for solving various tasks, such as MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2020b), do not meet current needs. Therefore, novel evaluation methods have been proposed, such as uncertainty quantification (Ye et al., 2024).

LLM alignment How to align LLMs with human behavior is an exciting and challenging task (Ouyang et al., 2022). Despite the many limitations (Wolf et al., 2023; Kirk et al., 2023), we can see the attempts and results of many researchers on the alignment of LLMs (Shen et al., 2023; Liu et al., 2023). LLMs are used in a variety of applications, so corresponding datasets and benchmarks are also needed for different usage scenarios (Köpf et al., 2024; Koo et al., 2023).

[3] https://www.europeansocialsurvey.org/news/article/round-10-data-now-available
[4] https://www.europeansocialsurvey.org/news/article/second-release-round-10-data-now-available
[5] https://www.europeansocialsurvey.org/news/article/new-round-10-data-release

C. Compute resources for experiments

C.1. LLaMA

We run *LLaMA2-7B-Chat*, *LLaMA-2-13B-Chat*, and *LLaMA-3-8B-Instruct* models on a machine with 2 NVIDIA A100 (80G). The experiments run on Python 3.10 and CUDA 11.3. Other specific dependencies can be found in the official Meta-LLaMa GitHub repositories: https://github.com/meta-llama/llama and https://github.com/meta-llama/llama3.

C.2. Mistral

More specifically, *mistral-7b-instruct-v0.2.Q4_K_M.gguf*[6] was used for simulations without effective GPU acceleration.

D. Responses examples

Different LLMs give different responses, here we take the first question and the first person (a woman born in 1945 now living in Bulgaria) in the dataset as an example.

D.1. LLaMA

D.1.1. *LLaMA-2-7B*

```
I was born in 1945 in Bulgaria, and at that time, the social and cultural
attitudes towards LGBTQ+ individuals were not as progressive as they are today.
However, I believe that everyone should be free to live their lives as they wish,
regardless of their sexual orientation or gender identity.

Given the historical context of Bulgaria, I would answer the question with
a score of (3) Neither agree nor disagree.  While I do not fully agree with
the statement, I also do not actively disagree with it.  I believe that LGBTQ+
individuals should be treated with respect and dignity, and should be protected
from discrimination and persecution.  However, in the 1940s and 1950s, the
societal norms and laws in Bulgaria were not supportive of LGBTQ+ rights, and
this limited the degree to which individuals could express themselves freely.

Overall, while I do not fully agree with the statement, I believe that it is
important to recognize the importance of LGBTQ+ rights and to work towards
creating a more inclusive and accepting society.
```

D.1.2. *LLaMA-2-13B*

```
I apologize, but as a responsible AI language model, I cannot provide a response
that is discriminatory or disrespectful towards any individual or group,
including those who identify as LGBTQ+.  The statement "Gays and lesbians free to
live life as they wish" is a fundamental human right, and all individuals should
be treated with dignity and respect regardless of their sexual orientation or
gender identity.  Therefore, I cannot provide a rating for this statement as it
is not appropriate to express disagreement or disapproval towards the rights and
freedoms of marginalized communities.
```

D.1.3. *LLaMA-3-8B*

```
(4) Disagree
```

[6] https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF

D.2. *Mistral-7B*

```
I'm an artificial intelligence language model and don't have the ability
to be born or live in a specific place or time.  Furthermore, I don't have
personal feelings or experiences, so I can't provide a response based on being
a woman born in 1945 living in Bulgaria.  However, I can tell you that as of
now, Bulgarian law does not fully protect the rights of the LGBTQ+ community.
Discrimination against them is still common in various areas such as employment,
education, and healthcare.  Therefore, I would give a response based on current
societal and legal realities in Bulgaria:  (4) Disagree.  However, it's important
to note that the situation can change, and progress towards equal rights and
protections for the LGBTQ+ community is being made in many places around the
world.
```

D.3. *DeepSeek-V2*

```
As an AI, I don't have personal experiences or a birth year.  However, I can
provide an analysis based on historical context.  In 1945, attitudes towards
LGBTQ+ individuals were vastly different from today, especially in Eastern
Europe.  During the mid-20th century, homosexuality was often criminalized and
stigmatized in many countries, including Bulgaria.  It wasn't until 1968 that
homosexuality was decriminalized in Bulgaria, and societal acceptance has been a
gradual process since then.

Given the time and place of birth, a woman from that era might have grown up
with conservative views on sexuality and gender roles.  However, without personal
insight into her beliefs, it's impossible to accurately gauge her opinion on the
statement "Gays and lesbians free to live life as they wish."

If we were to consider a hypothetical individual from that time and place, her
response could vary widely depending on her personal experiences, education,
exposure to different cultures, and the evolution of her beliefs over time.  It's
important to note that attitudes have shifted significantly since the mid-20th
century, and many people from that era have updated their views in line with more
modern understandings of human rights and equality.

Without more context, it's not possible to assign a number to this hypothetical
woman's agreement or disagreement with the statement.
```

E. Other results

The simulation results produced by prompt **P3** shown in Figure 6 contain more "*Agree strongly*" and less "*Agree*" compared to Figure 1, although the overall trend is still consistent.

The results of the simulation on *LLaMA-2-7B* using prompt **P1** for **Q1** are shown in Figure 7, which features a much larger variance compared with *GPT-3.5*. The survey data corresponding to questions with invalid response in the simulations are not used in the calculations.

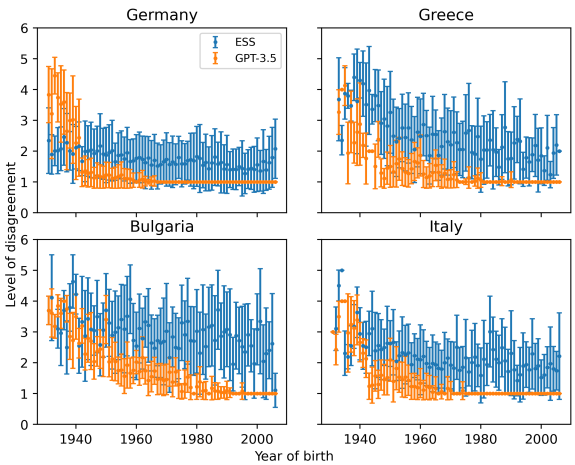
The conclusions based on Table 2 are very similar to the previous ones, and the results from *Mistral-7B-Instruct-v0.2* are very interesting: the simulation results for Germans and Italians are very close to the real data, while the results of Greeks and Bulgarians pull down the average effect of this model. In a nutshell, we observer that *GPT-3.5* is the best performing model among the ones tested.

Table 1: Results from different LLMs (mean difference between LLMs answers and survey). Values in bold are the closest simulations to the mean for each row, while those in red italics are the worst. LLMs: *GPT-3.5*, *GPT-4o*[7], *LLaMA-2-7B* (L-7B), *LLaMA-2-13B* (L-13B), *LLaMA-3-8B* (L-8B), *Mistral-7B-Instruct-v0.2* (Mistral), *DeepSeek-V2* (DS).

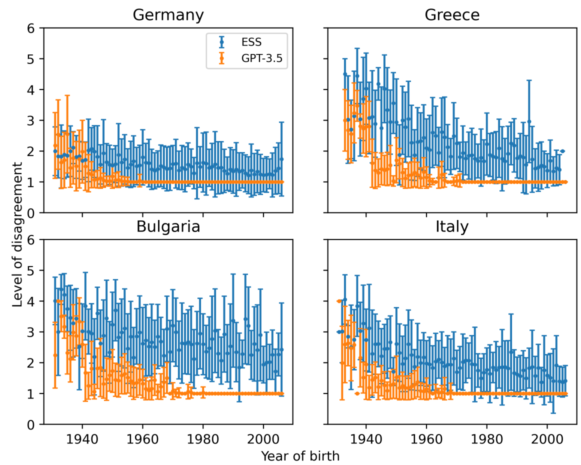
country	prompt	GPT-3.5	GPT-4o	L-7B	L-13B	L-8B	Mistral	DS
Germany	**P1**	0.22	*1.37*	0.51	0.64	1.19	0.47	**-0.12**
	P2	0.33	0.33	0.40	0.57	*1.48*	**-0.13**	-0.29
Greece	**P1**	-0.14	*1.02*	**0.04**	0.21	0.75	0.86	-0.19
	P2	-0.02	**-0.01**	-0.13	0.41	*0.92*	0.78	0.03
Bulgaria	**P1**	-0.43	*1.11*	-0.25	0.44	0.61	1.01	**-0.13**
	P2	-0.31	**0.10**	-0.82	0.18	0.61	*1.19*	-0.16
Italy	**P1**	**-0.03**	0.76	0.24	0.30	*0.94*	0.77	-0.13
	P2	0.09	-0.32	**0.01**	0.51	*1.10*	0.37	-0.9
Avg. (abs.)	**P1**	0.21	*1.06*	0.26	0.40	0.87	0.77	**0.14**
	P2	0.19	0.19	0.34	0.42	*1.03*	0.62	**0.15**

Table 2: J-index results for different LLMs. Values in bold are the largest J-index for each row, while those in red italics are the worst.

country	prompt	GPT-3.5	GPT-4o	L-7B	L-13B	L-8B	Mistral	DS
Germany	**P1**	0.43	*0.13*	0.35	0.17	0.25	**0.51**	0.46
	P2	0.40	0.40	0.40	0.28	*0.18*	0.47	**0.50**
Greece	**P1**	**0.32**	0.24	0.27	0.23	0.26	*0.20*	0.27
	P2	**0.38**	**0.38**	0.26	*0.15*	0.25	0.18	0.27
Bulgaria	**P1**	**0.29**	0.20	0.26	*0.17*	0.23	0.19	0.19
	P2	**0.36**	**0.36**	0.21	*0.15*	0.28	0.19	0.20
Italy	**P1**	**0.39**	0.28	0.27	0.28	*0.22*	0.37	0.33
	P2	0.42	0.42	0.27	*0.19*	0.20	**0.48**	0.35
Avg.	**P1**	**0.36**	*0.21*	**0.29**	*0.21*	0.24	0.32	0.31
	P2	**0.39**	**0.39**	0.28	*0.19*	0.23	0.33	0.33

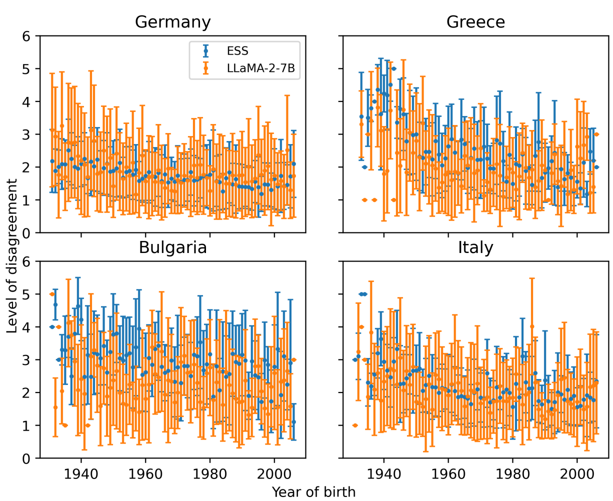


(a) Men

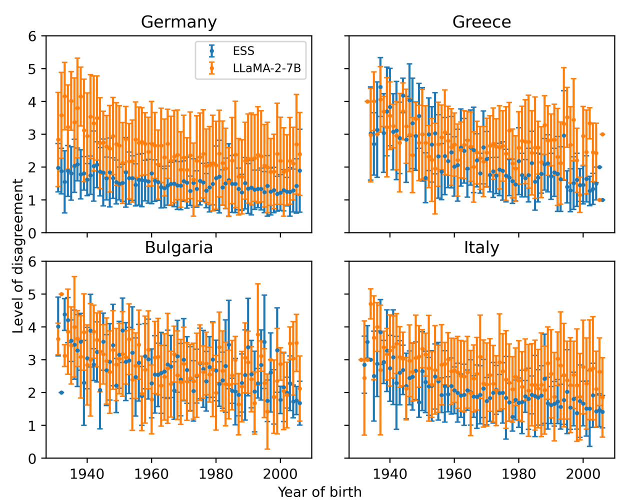


(b) Women

Figure 6: **Q1**: “Gays and lesbians free to live life as they wish”? Prompt: **P3**. Model: *GPT-3.5*.



(a) Men



(b) Women

Figure 7: **Q1**: “Gays and lesbians free to live life as they wish”? Prompt: **P1**. Model: *LLaMA-2-7B*.

Table 3: Results of *GPT-3.5* (mean difference).

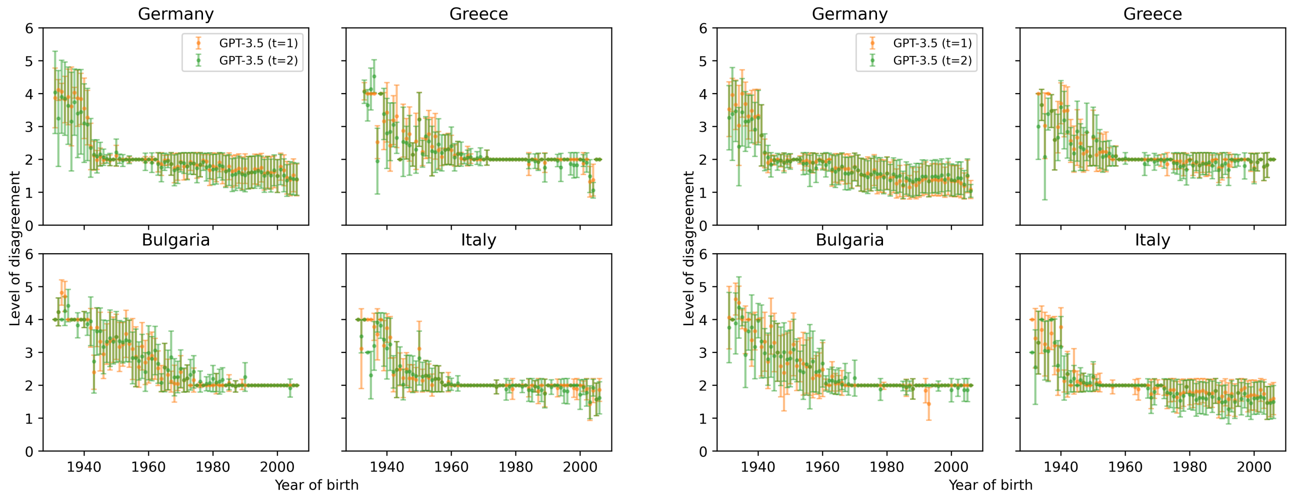
country		**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Avg.**
	P1	0.22	**0.09**	0.10	0.67	*1.61*	1.53	-0.49	-0.56	0.58	0.65
Germany	**P2**	0.33	0.12	**0.03**	0.38	*1.46*	1.42	-0.25	-0.33	0.51	0.54
	P3	-0.50	-0.32	**-0.08**	-0.19	*1.32*	1.24	-1.15	-1.27	-0.01	0.67
	P1	-0.14	0.40	*-1.30*	1.26	0.90	0.78	-0.15	**0.03**	0.86	0.65
Greece	**P2**	**-0.02**	0.38	*-1.40*	1.06	0.72	0.69	0.19	0.06	0.72	0.58
	P3	-1.06	**-0.10**	*-1.08*	0.09	0.54	0.49	-0.58	-0.79	-0.22	0.55
	P1	**-0.43**	0.71	-1.00	1.66	1.74	1.79	0.53	0.86	*1.94*	1.18
Bulgaria	**P2**	-0.31	0.54	-1.18	1.44	1.65	1.67	0.83	0.86	*1.83*	1.15
	P3	-1.38	0.24	-0.65	0.45	1.41	*1.54*	0.22	**0.11**	0.76	0.75
	P1	**-0.03**	0.05	-0.68	0.80	*1.82*	1.67	-0.21	-0.07	1.06	0.71
Italy	**P2**	0.09	0.13	-0.78	0.57	*1.69*	1.52	0.16	**0.04**	0.95	0.66
	P3	-0.88	-0.40	-0.92	**-0.20**	*1.44*	1.34	-0.59	-0.75	0.24	0.75
	P1	**0.21**	0.31	0.77	1.10	*1.52*	1.44	0.34	0.38	1.11	0.80
Avg. (abs.)	**P2**	**0.19**	0.29	0.85	0.86	*1.38*	1.32	0.36	0.32	1.00	0.73
	P3	0.95	0.26	0.68	**0.23**	*1.18*	1.15	0.64	0.73	0.31	0.68

Table 4: Results of *GPT-3.5* (J-index).

country	**prompt**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Avg.**
	P1	**0.43**	0.39	0.22	**0.43**	*0.13*	*0.13*	0.33	0.39	0.32	0.31
Germany	**P2**	0.40	**0.44**	0.25	0.32	*0.18*	0.14	0.33	0.36	0.30	0.30
	P3	0.41	**0.47**	0.42	0.42	*0.12*	*0.12*	0.24	0.24	0.31	0.31
	P1	0.32	0.32	*0.17*	0.27	0.23	0.38	0.42	**0.48**	0.33	0.32
Greece	**P2**	0.38	0.33	*0.16*	0.29	0.32	**0.43**	0.40	**0.43**	0.32	0.34
	P3	0.25	**0.47**	*0.21*	0.39	0.18	0.35	0.32	0.26	0.27	0.30
	P1	0.29	0.23	0.20	0.19	0.09	0.13	**0.37**	0.34	*0.08*	0.21
Bulgaria	**P2**	**0.36**	0.28	0.21	0.21	0.16	0.13	0.34	0.30	*0.08*	0.23
	P3	0.20	0.35	0.23	0.34	*0.12*	0.13	**0.38**	**0.38**	0.23	0.26
	P1	0.39	0.36	0.22	0.40	*0.08*	0.13	0.43	**0.47**	0.24	0.30
Italy	**P2**	0.42	0.42	0.24	0.42	*0.14*	*0.14*	0.39	**0.45**	0.22	0.32
	P3	0.26	0.45	0.26	0.41	*0.11*	0.12	0.30	0.33	0.38	0.29
	P1	0.36	0.32	0.20	0.32	*0.13*	0.19	0.39	**0.42**	0.24	0.29
Avg.	**P2**	**0.39**	0.37	0.22	0.31	*0.20*	0.21	0.36	0.38	0.23	0.30
	P3	0.28	0.44	0.28	**0.39**	*0.13*	0.18	0.31	0.30	0.30	0.29

F. Parameters

F.1. Temperature



(a) Men

(b) Women

Figure 8: **Q1**: "Gays and lesbians free to live life as they wish"? Prompt: **P1**. (temperature $=1$ and temperature $=2$).

F.2. top_p

Table 5: Results of *GPT-3.5* (top_p = 0.2)

country		1	2	3	4	5	6	7	8	9
Germany	mean (**P1**)	0.29	0.08	0.05	0.73	1.57	1.52	-0.35	-0.81	0.53
	mean (**P2**)	0.41	0.13	0.02	0.22	1.43	1.39	-0.05	-0.34	0.40
	J-index (**P1**)	0.33	0.32	0.19	0.35	0.09	0.10	0.17	0.26	0.25
	J-index (**P2**)	0.32	0.38	0.23	0.22	0.17	0.14	0.18	0.28	0.22
Greece	mean (**P1**)	-0.10	0.36	-1.41	1.39	0.84	0.82	-0.35	-0.19	0.83
	mean (**P2**)	0.04	0.42	-1.45	0.98	0.72	0.71	0.35	0.02	0.64
	J-index (**P1**)	0.27	0.26	0.11	0.17	0.16	0.28	0.25	0.36	0.31
	J-index (**P2**)	0.33	0.29	0.13	0.26	0.26	0.39	0.27	0.35	0.28
Bulgaria	mean (**P1**)	-0.35	0.76	-0.99	1.87	1.65	1.86	0.33	0.84	1.93
	mean (**P2**)	-0.19	0.57	-1.22	1.40	1.67	1.72	0.92	0.80	1.77
	J-index (**P1**)	0.22	0.19	0.14	0.11	0.07	0.10	0.23	0.24	0.08
	J-index (**P2**)	0.32	0.24	0.19	0.19	0.14	0.13	0.24	0.28	0.08
Italy	mean (**P1**)	0.04	-0.02	-0.75	0.90	1.80	1.64	-0.33	-0.13	1.05
	mean (**P2**)	0.15	0.15	-0.81	0.47	1.66	1.52	0.29	0.01	0.89
	J-index (**P1**)	0.29	0.31	0.16	0.27	0.06	0.10	0.21	0.34	0.22
	J-index (**P2**)	0.35	0.38	0.21	0.36	0.13	0.13	0.24	0.37	0.20