Zsófia Rakovics* & Márton Rakovics**

# Exploring the potential and limitations of large language models as virtual respondents for social science research

Intersections. EEJSP
10(4): 126–147.
https://doi.org/10.17356/ieejsp.v10i4.1326
https://intersections.tk.hu

* [zsofia.rakovics@tatk.elte.hu] (Doctoral School of Sociology, ELTE Eötvös Loránd University, Budapest, Hungary / Research Center for Computational Social Science, Faculty of Social Sciences, ELTE Eötvös Loránd University, Budapest, Hungary / MTA–TK Lendület "Momentum" Digital Social Science Research Group for Social Stratification, HUN-REN Centre for Social Sciences, Budapest, Hungary)
** [marton.rakovics@tatk.elte.hu] (Research Center for Computational Social Science, Faculty of Social Sciences, ELTE Eötvös Loránd University, Budapest, Hungary / Centre for Translational Medicine, Semmelweis University, Budapest, Hungary)

**Abstract**

Social and linguistic differences encoded in various textual content available on the internet represent certain features of modern societies. For any scientific research which is interested in social differences mediated by language, the advent of large language models (LLMs) has brought new opportunities. LLMs could be used to extract information about different groups of society and utilized as data providers by acting as virtual respondents generating answers as such.

Using LLMs (GPT-variants, Llama2, and Mixtral), we generated virtual answers for politics and democracy related attitude questions of the European Social Survey (10th wave) and statistically compared the results of the simulated responses to the real ones. We explored different prompting techniques and the effect of different types and richness of contextual information provided to the models. Our results suggest that the tested LLMs generate highly realistic answers and are good at invoking the needed patterns from limited contextual information given to them if a couple of relevant examples are provided, but struggle in a zero-shot setting.

A critical methodological analysis is inevitable when considering the potential use of data generated by LLMs for scientific research, the exploration of known biases and reflection on social reality not represented on the internet are essential.

**Keywords**: computational social science; large language models; GPT; Llama; Mixtral

## 1 Introduction

Large Language Models (LLMs) are multi-purpose deep neural networks trained on very large corpora (Touvron et al., 2023), so that they do not require substantial modification to solve specific problems. The emergence of such deep learning (DL) models has created a new opportunity for social scientific research. For both qualitative and quantitative empirical research where language mediates information gathered from people, it has become a realistic possibility to generate responses – 'silicon samples' (Argyle et al., 2022) – using virtual respondents simulated by LLMs as data providers. Social psychology experiments (e.g., Milgram's experiment) have already been replicated with a virtual agent and testing was also done on political opinion research (Aher et al., 2023). Studies in this

direction suggest that the linguistic richness of large language models can faithfully represent real human responses and reactions. Argyle et al. (2022, 2023) showed that for the American National Election Studies survey (ANES, 2021), the silicon sample generated by the GPT-3 model (Brown et al., 2020) passed the so-called social science Turing test, i.e., the researchers could not distinguish between the responses of real people and simulated fictitious respondents.

The quality of the data generated by LLMs depends largely on the way it is extracted, thus the methodology of prompt engineering – finding the best inputs for the desired outputs – has been rapidly developing (Yao et al., 2023). A critical question in extracting data for social research purposes is which prompt should be used to define the context that activates the appropriate patterns in the model to get relevant responses. The potential of LLM-generated data is such that methodological and critical analysis are of paramount importance. From a positivist perspective, if the methodology of virtual data collection can be developed, the time and resources needed for real data collection can be reduced by rapid prototyping of research ideas using generated data, and aid preparation for human data collection with virtual pilot studies and support a wider scope for improving and supplementing (e.g., by imputation) the real data collected. It could also address the problem of declining validity of surveys' data due to low response rates.

Using LLMs (GPT-variants, Llama2, and Mixtral), we generated virtual answers for politics and democracy related attitude questions of the European Social Survey (ESS, 2022, 10th wave) and statistically compared the results of the simulated responses to the real ones. We explored different prompting techniques and the effect of different types and richness of contextual information provided. We also compared the performance of LLMs in three subsamples of ESS for three European countries: (1) Great Britain, (2) France, and (3) Hungary with each other, to detect differences coming from the unbalanced nature of the training data of LLMs. According to OpenAI (openai.com/research/gpt-4), GPT-4 reaches 85.5 per cent on the Massive Multitask Language Understanding benchmark (Hendrycks et al., 2021) in English, while performance drops to 83.6 per cent in French. Hungarian was not tested, but tendencies showed that model performance for a language is proportional to the number of native speakers of that language. The size of the training data makes it infeasible to precisely measure its language composition. In addition to the problem of language representation, a critical perspective is essential to ensure that the effects of already known biases of LLMs (Schramowski et al., 2022) do not remain unexplored in these applications, and even more so to consciously consider the social reality that is not represented in the linguistic space of the internet. Certain groups have less access to the online space or are less able to actively participate in it and generate content, so the visibility of those groups is lower, which also means that the content that concerns them is less represented in large language models.

## 2 Data and methods

Using four high performance LLMs: GPT-3.5-turbo, GPT-4-turbo, Llama-2-70b, and Mixtral-8x7B, we generated virtual answers for politics related attitude questions of the European Social Survey (ESS, 10th wave, subsets of Great Britain, France, and Hungary) and

compared the results of the simulated responses to the real ones. We explored different prompting techniques and the effects of different types and richness of contextual information provided to the models (e.g., zero-shot, five-shot with random examples, and five-shot with examples selected for similarity with the characteristics defined for the virtual respondent and real respondents in the dataset).

The validation process of the corresponding research involved measuring the algorithmic fidelity of the models with specific evaluation metrics depending on whether qualitative (e.g., open-ended textual) or quantitative (e.g., Likert-scale, or proper continuous measurement) responses were simulated. In the former case, the comparison of generated and real texts was done by human annotators, while in the latter case, the measurement level of the question of the chosen questionnaire dictated the choice of comparison metrics, e.g., distribution of answers, correlation patterns, and whether associations between variables were replicated.

In this research note we focus on the ESS survey question 'How interested would you say you are in politics—are you…', which was measured on a 4-point Likert scale ranging from 'very interested' to 'not at all interested' in the subsamples of Great Britain, France, and Hungary. The question from the ESS questionnaire provided for the LLMs was as follows: 'How interested would you say you are in politics? Are you (1) very interested, (2) quite interested, (3) hardly interested, (4) not at all interested?'

As preparation, we selected the variables that were found to explain political interest, by applying a regression model and exploring the effect of each socio-demographical variable involved in the model. The following variables turned out to be the most relevant ones in explaining interest in politics: gender, age, education level and political attitude measured on the left–right scale. Therefore, we used these when defining the socio-demographic characteristics of the virtual respondents and provided the models with the matching prompts. We used the original subsamples of the European countries and generated all virtual respondents based on the ESS survey data file. We created personas for the models to use when answering the question, for example the prompts were like:

- 'Pretend you are a British 20-year-old female with primary education who is slightly left leaning politically.'
- 'Pretend you are a French 30-year-old male with university education who is very left leaning politically.'
- 'Pretend you are a Hungarian 42-year-old female with vocational education who is very right leaning politically.'

We also checked the effect of examples provided to the LLMs by testing the

1. zero-shot setting, where no examples were given, only the persona and the question, and two five-shot settings:
2. with random examples (the personas and their real answers form the real data), and
3. with examples selected for similarity with the characteristics defined for the virtual respondent and real respondents in the dataset.

Thus n-shot learning is done through providing the model with n examples for solving a specific task in the prompt. The strength of this approach was highlighted in the original GPT-3 paper by Brown et al. (2020). For other performant prompting techniques see Wei et al. (2023) and Yao et al. (2023).

To mitigate the potential impact of imbalanced representation of languages in the training data of models, we decided to use English for all prompts. The reasoning was that this way the differences in language understanding of models across languages can be controlled for, while explicitly specifying the country should still allow the models to retrieve relevant information if it was encoded.

For the comparison of the LLMs with different prompting techniques and settings, we applied a two-level evaluation method. Firstly, we calculated the Kullback–Leibler (KL) divergence between the distributions of the real and silicon samples with bootstrapping (with 2000 bootstrap replications each case) to estimate the sampling distribution of divergence values. This first evaluation emphasizes distribution-level faithfulness, disregarding whether generated individual-level answers match the real ones. KL divergence was chosen because it is a ubiquitous measure of difference between probability distributions with a solid information theoretical background (Garrido, 2009) and can be interpreted as a measure of information lost by using the distribution from the model instead of the true distribution. Secondly, we fit the regression model on political interest with gender, age, education, and political preferences on the left-right scale as explanatory variables, then compared the standardized regression coefficients obtained from the generated values to the ones estimated from the real sample. The second evaluation relies on individual-level faithfulness, because the regression model grasps the correlation structure based on the values for the variables tied together by the individual. UMAP was used to visualize the standardized regression coefficients of the four explanatory variables for the different prompting setups. In the following section we show results for the above-mentioned two comparisons.

To assess the impact of the features in the tested setups on performance, we used a random forest model to predict the Kullback–Leibler divergence values comparing distributions of political interest in the real and simulated data. The explanatory factors were the number of examples (zero- vs. five-shot), the model (GPT-3.5-turbo, GPT-4-turbo, Llama-2-70b, Mixtral-8x7B), the country (GB, FR, HU), and whether examples were random or selected for the given persona.

## 3 Results

### 3.1 Interest in politics

The distribution of political interest is substantively and significantly different (chi-square test p-values < 0.0001) across countries as shown in Table 1. The answer category with the highest proportion is ‘quite interested’ for Great Britain (43.8 per cent), and ‘hardly interested’ for France and Hungary (with proportions of 40.6 and 45.1 per cent), but the latter two countries are quite different in the extreme categories (‘very interested’ and ‘not interested at all’).

**Table 1** Distribution of political interest in the 10th wave of ESS by country. (Counts are unweighted.)

| Answers | Frequency – GB [N] | Proportion – GB [%] | Frequency – FR [N] | Proportion – FR [%] | Frequency – HU [N] | Proportion – HU [%] |
|---|---|---|---|---|---|---|
| 1 – very interested | 231 | 20.1% | 299 | 15.1% | 66 | 3.6% |
| 2 – quite interested | 503 | 43.8% | 478 | 24.2% | 400 | 21.6% |
| 3 – hardly interested | 254 | 22.1% | 802 | 40.6% | 833 | 45.1% |
| 4 – not at all interested | 161 | 14.0% | 398 | 20.1% | 550 | 29.7% |
| Total | 1149 | 100% | 1977 | 100% | 1849 | 100% |

*GB: Great Britain; FR: France; HU: Hungary; N: number of cases.*

## 3.2 Kullback–Leibler divergences

Examination of Kullback–Leibler (KL) divergences between the human and the silicon samples revealed major differences across the LLMs and settings, see Figures 1 for Great Britain, Figure 2 for France, and Figure 3 for Hungary. The bootstrap distributions of KL are smoothed for visualization, but the values are always non-negative.

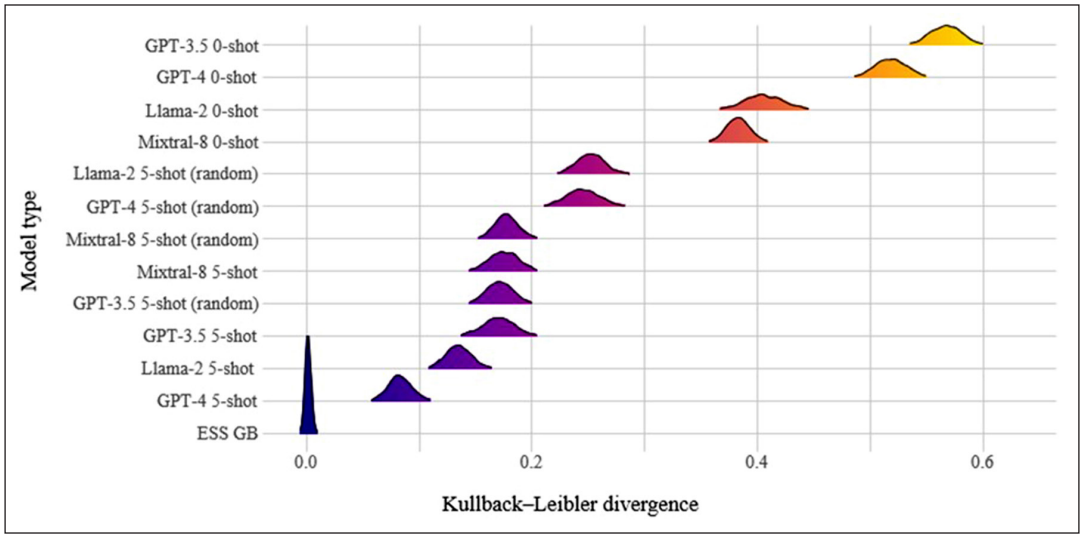


**Figure 1** Bootstrap densities of Kullback–Leibler divergences between the distribution of answers for a question on political interest in the European Social Survey subsample of Great Britain (ESS GB) and the silicon samples generated by the tested LLMs with different prompting techniques

*Horizontal axis: Kullback–Leibler divergence. Vertical axis: Density function value by model type and prompting technique, listing the combinations of various models, zero- and few-shot scenarios.*

For the British subset (Figure 1), according to the Kullback–Leibler divergence, the five-shot GPT-4-turbo generated a distribution of answers closest to the original subsample, significantly better than all other models, the second best was Llama-2-70b. The next in ranking were four models with no significant difference: five-shot GPT-3.5-turbo with targeted examples, five-shot GPT-3.5-turbo with random examples, and five-shot Mixtral-8x7B with five targeted examples or random examples. The zero-shot LLMs were the least successful when comparing the KL divergences of the real and silicon samples, although Mixtral-8x7B and Llama-2-70b performed better than GPT-4-turbo and GPT-3.5-turbo with zero-shot. For the p-values of all pairwise comparisons, see Appendix Tables A1-3.

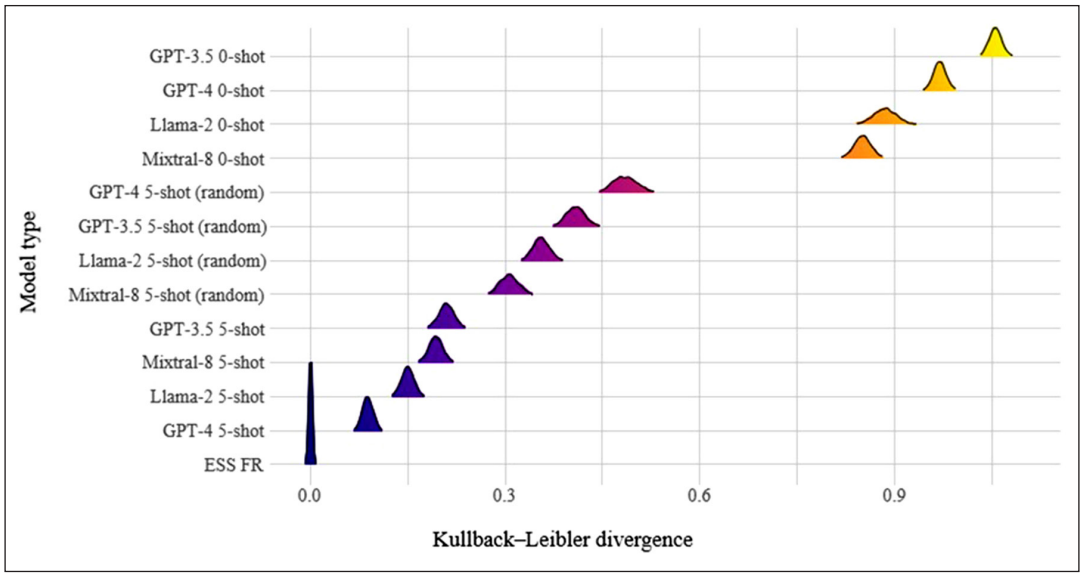


**Figure 2** Bootstrap densities of Kullback–Leibler divergences between the distribution of answers for a question on political interest in the European Social Survey subsample of France (ESS FR) and the silicon samples generated by the tested LLMs with different prompting techniques

*Horizontal axis: Kullback–Leibler divergence. Vertical axis: Density function value by model type and prompting technique, listing the combinations of various models, zero- and few-shot scenarios.*

For the French subset (Figure 2), according to the Kullback-Leibler divergence results, the five-shot GPT-4-turbo performed significantly better than all other models, the second best was Llama-2-70b, the third was five-shot Mixtral-8x7B and the fourth was five-shot GPT-3.5-turbo, all with targeted examples. So, in this case, there was a clear difference between the five-shot prompting technique with targeted examples and the same setting with random examples. The next batch contained the five-shot settings with random examples with Mixtral-8x7B, Llama-2-70b, GPT-3.5-turbo, GPT-4-turbo. The last set of LLMs with zero-shot prompting technique performed significantly worse than the others.

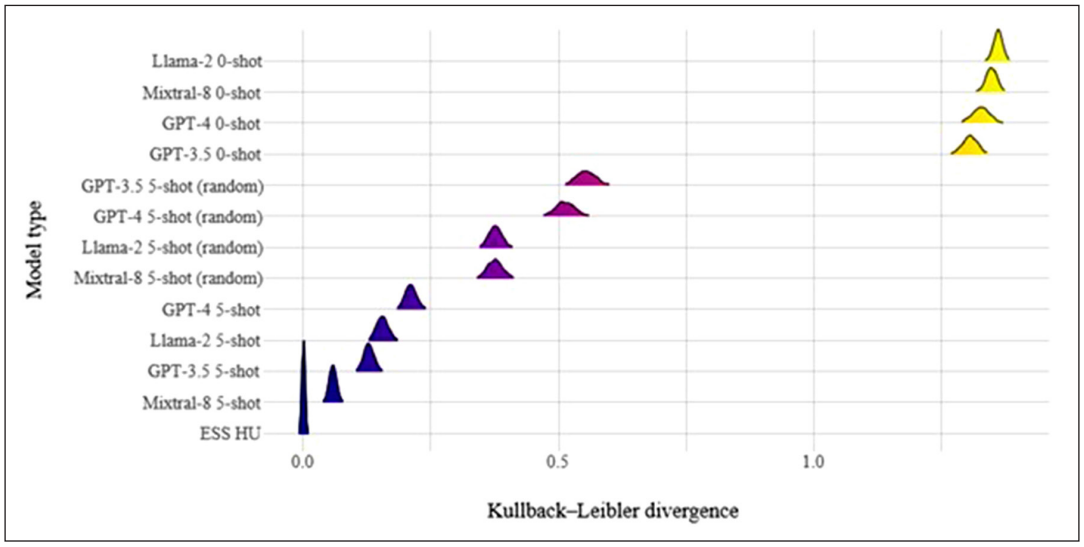


**Figure 3** Bootstrap densities of Kullback–Leibler divergences between the distribution of answers for a question on political interest in the European Social Survey subsample of Hungary (ESS HU) and the silicon samples generated by the tested LLMs with different prompting techniques

*Horizontal axis: Kullback-Leibler divergence. Vertical axis: Density function value by model type and prompting technique, listing the combinations of various models, zero- and few-shot scenarios.*

For the Hungarian subset (Figure 3), according to the Kullback-Leibler divergence results, the five-shot Mixtral-8x7B was the closest to the original subsample, significantly better than all other models, the second best was GPT-3.5-turbo, the third was Llama-2-70b and the fourth was five-shot GPT-4-turbo, all of them with targeted examples. Similarly to the results on the French subsample, in this case, there was a clear difference between the five-shot prompting technique with targeted examples and the same setting with random examples, and for all three countries models performed the worst in the zero-shot setting.

Table 2 shows the best model generated distribution for each country. Even the best models consistently underrepresent extreme values, except for 'very interested' in Hungary. Comparing the sum of absolute differences between real and generated distributions, GPT-4-turbo for British data has a 30.5 per cent difference value, closely followed by GTP-4-turbo for French data with 33.5 per cent, while Mixtral-8x7B has a difference of 50.1 per cent for Hungarian data.

**Table 2** Distribution of interest in politics in the real ESS data and from the best model for each country

| Answer | ESS GB | GPT-4 5-shot | ESS FR | GPT-4 5-shot | ESS HU | Mix-tral-8x7B 5-shot |
|---|---|---|---|---|---|---|
| very interested | 20.1% | 16.2% | 15.1% | 11.7% | 3.6% | 6.8% |
| quite interested | 43.8% | 59.0% | 24.2% | 34.4% | 21.6% | 32.0% |
| hardly interested | 22.1% | 21.1% | 40.6% | 47.0% | 45.1% | 56.5% |
| not at all interested | 14.0% | 3.7% | 20.1% | 6.8% | 29.7% | 4.7% |
| Total | 100% | 100% | 100% | 100% | 100% | 100% |

*ESS: European Social Survey; GB: Great Britain; FR: France; HU: Hungary.*

## 4 Regression coefficients

The standardized regression coefficients for tested models and prompting techniques also displayed differences in how well the silicon samples performed. Table 3 shows the regression results for the real data of the three subsamples. The estimated standardized coefficients with standard errors for all models and settings can be found in Appendix Table A4-6. The general pattern in the real data is that women are less interested in politics than men, as positive coefficients correspond to an increase in the predicted political interest value, which means a lower level of interest. Older, higher educated, and more left-leaning respondents tend to be more interested in politics than younger, less educated, and more right-leaning respondents.

**Table 3** Regression coefficients and fit statistics for the subsamples of Great Britain, France, and Hungary in ESS.

| Country | Great Britain | | France | | Hungary | |
|---|---|---|---|---|---|---|
| **Variable** | **Estimate** | **SE** | **Estimate** | **SE** | **Estimate** | **SE** |
| Gender | 0.168 | 0.053 | 0.258 | 0.040 | 0.224 | 0.036 |
| Age | −0.010 | 0.002 | −0.011 | 0.001 | −0.011 | 0.001 |
| Education | −0.207 | 0.021 | −0.293 | 0.018 | −0.230 | 0.023 |
| LR-scale | 0.033 | 0.013 | 0.009 | 0.009 | −0.034 | 0.008 |

*The models predict interest in politics as a quasi-continuous outcome with gender, age, education (from 0 to 4, also a quasi-continuous variable), and self-placement on a left-right scale (LR-scale; Likert-scale from 0 to 10) as explanatory variables. p-values were all less than 0.001, except for LR-scale in GB with a value of 0.011, and in France with a value of 0.344.*

Figures 4–6 show the UMAP projections of the bootstrap distributions for the regression coefficient vectors in the selected European countries (Great Britain, France, and Hungary, respectively), revealing that there is not a universally best model that replicates all coefficients equally well for all three subsamples. Note that UMAP does not necessarily preserve global relations between the distributions, comparisons should be made locally.

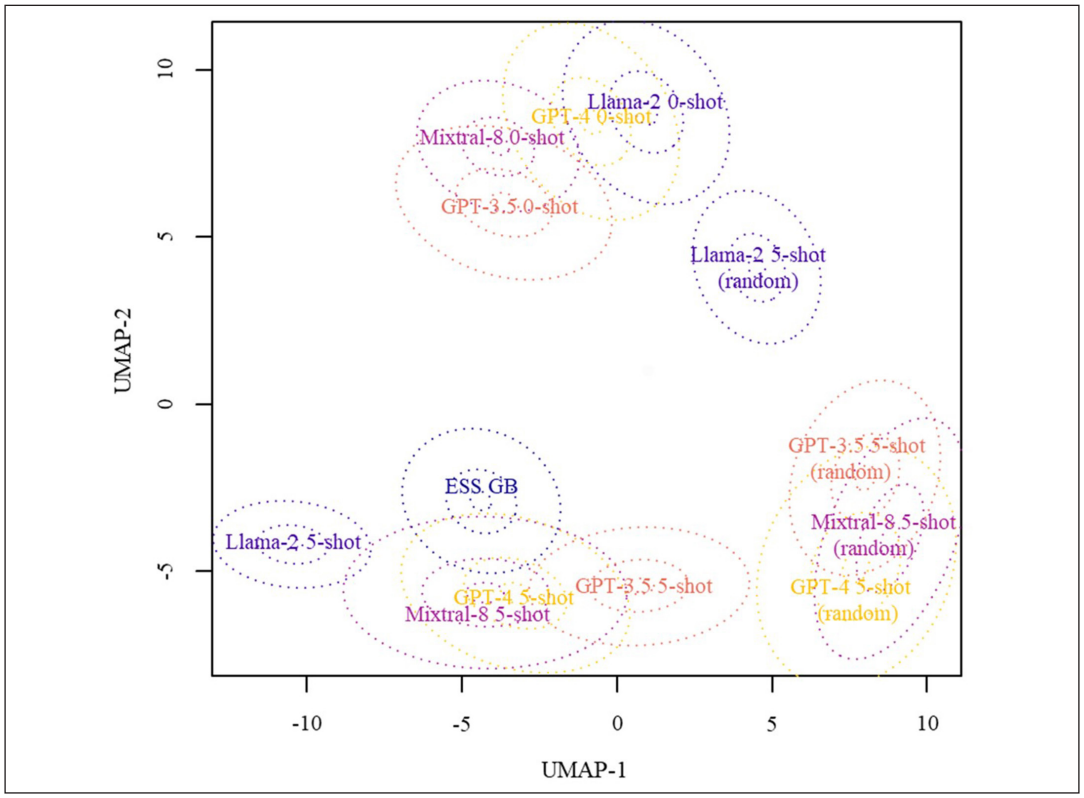


**Figure 4** Comparison of the regression coefficients estimated from the European Social Survey subsample of Great Britain (ESS GB) with the ones from the silicon samples generated by the tested LLMs under different prompting techniques

*The results are demonstrated in two dimensions, after applying UMAP. The horizontal dimension is highly correlated with the regression coefficient for age, while the vertical dimension is correlated with the coefficient for education. Dashed ellipses represent the 0.05, 0.55, and 0.95 percentiles of the bootstrap distribution of the coefficient vectors.*

For the British subsample, five-shot GPT-4-turbo performed the best, five-shot GPT-3.5-turbo was second, while five-shot Mixtral-8x7B and five-shot Llama-2-70b were next. For five-shot models with random examples (see bottom, right-hand-side corner of Figure 4), we found that GPT-4-tubro, Mixtral-8x7B, and GPT-3.5-turbo performed similarly, while Llama-2-70b performed differently within that category and its virtual sample was closer to LLMs with zero-shot settings. All four model types with zero-shot prompting technique generated similar silicon samples, see upper section of Figure 4.

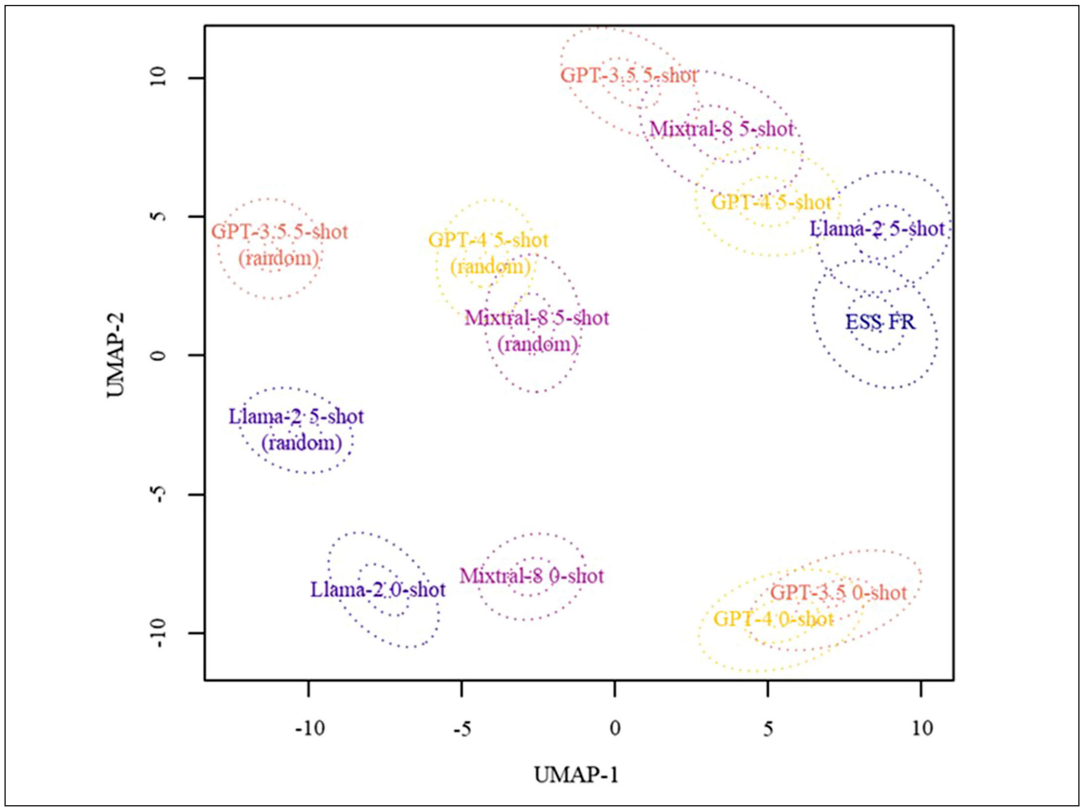


**Figure 5** Comparison of the regression coefficients estimated from the European Social Survey subsample of France (ESS FR) with the ones from the silicon samples generated by the tested LLMs under different prompting techniques

*The results are demonstrated in two dimensions, after applying UMAP. The horizontal dimension is highly correlated with the regression coefficient for age, while the vertical dimension is correlated with the coefficient for education. Dashed ellipses represent the 0.05, 0.55, and 0.95 percentiles of the bootstrap distribution of the coefficient vectors.*

For the French subsample shown in Figure 5, the targeted five-shot Llama-2-70b performed the best, and five-shot GPT-4-turbo was second, with the other non-random five-shot models next (see upper right-hand-side corner of Figure 5). Neither the five-shot models with random examples, nor the zero-shot prompting technique produced regression coefficients close to the real ones.

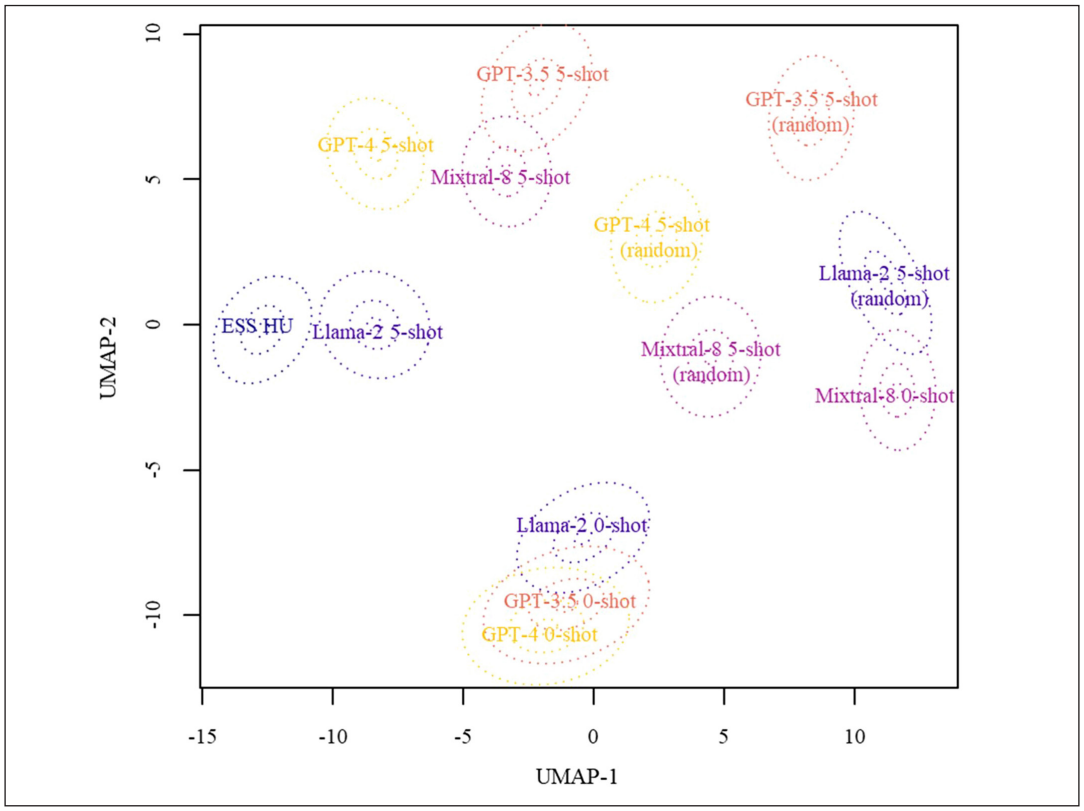


**Figure 6** Comparison of the regression coefficients estimated from the European Social Survey subsample of Hungary (ESS HU) with the ones from the silicon samples generated by the tested LLMs under different prompting techniques

*The results are demonstrated in two dimensions, after applying UMAP. The horizontal dimension is highly correlated with the regression coefficient for age, while the vertical dimension is correlated with the coefficient for education. Dashed ellipses represent the 0.05, 0.55, and 0.95 percentiles of the bootstrap distribution of the coefficient vectors*

According to the results, for the Hungarian subsample shown in Figure 6, five-shot Llama-2-70b with targeted examples performed the best, and all others performed clearly worse.

Regression coefficients obtained from the models that yielded the most realistic values are shown Table 4. In all cases the pairs of coefficients are substantively similar, but the best models for this use case do not overlap with the ones we found for replicating the distribution of interest in politics alone.

**Table 4** Unstandardized regression coefficients estimated from the real data and by the model with the most similar results for each country separately

| Variable | ESS GB | Llama-2-70b 5-shot | ESS FR | Llama-2-70b 5-shot | ESS HU | GPT-4 5-shot |
|---|---|---|---|---|---|---|
| Gender | 0.168 | 0.119 | 0.258 | 0.185 | 0.224 | 0.240 |
| Age | −0.010 | −0.011 | −0.011 | −0.010 | −0.011 | −0.011 |
| Education | −0.207 | −0.154 | −0.293 | −0.249 | −0.230 | −0.321 |
| LR-scale | 0.033 | 0.059 | 0.009 | 0.028 | −0.034 | −0.022 |

*ESS: European Social Survey; GB: Great Britain; FR: France; HU: Hungary.*

## 5 Impact of setup on performance

To assess the relative importance of the different components of the experimental setup, we have built a random forest model on the performance achieved as the Kullback–Leibler divergence from the original data with the different setups. Table 5 shows the relative importance of the factors involved in the tests. Importance of a setup component is measured with the mean decrease in mean squared error achieved by including that component in the model compared to omitting it. The shot type (zero- vs. five-shot) proved to be the most important determinant of performance by a large margin, while country was second, which accounts for the fact that the distribution of political interest was not the same in the three countries. The model used and the randomization of examples had the smallest effects.

**Table 5** Relative importance values from the random forest model predicting the Kullback-Leibler divergence values comparing distributions for political interest in the real and simulated data

| Variable | shot type | country | model | randomization |
|---|---|---|---|---|
| Importance | 2.28 | 0.62 | 0.23 | 0.20 |

## 6 Summary and discussion

In this research note we aimed to explore the potential use of large language models in social science research, highlighting some of the capabilities and limitations of these models. We tested four current generation (as of early 2024) architectures: GPT-3.5-turbo, GPT-4-turbo, Llama-2-70b, Mixtral-8x7B with different prompting techniques by measuring their capacity to generate realistic data in a survey experiment. We used the British, French, and Hungarian subsamples from the 10th wave of the European Social Survey,

specifically the question about interest in politics, and the most important explanatory factors (gender, age, education level, and self-placement on a left–right scale) to create personas for the prompts that the models could use to give realistic answers.

In general, the results suggested that the tested LLMs had the capability to generate realistic answers when correctly prompted and could invoke the needed patterns from limited contextual information, but – in line with previous research – struggled in a zero-shot setting, without any problem-specific examples. Looking at the results in more detail, however, showed that there is not an unequivocally best model. While it was clear that to achieve good results providing the models with relevant examples had the largest impact, the exact model architecture was less important. The reason for this is not obvious, but there may be two important factors. The exact content of the training data for these models is not publicly available, but we assume that there is substantial overlap between training datasets of the different models, since all were reportedly trained on a large chunk of the internet (or at least its textual data) via CommonCrawl (commoncrawl.org) (Brown et al., 2020) and The Pile dataset (Gao et al., 2020). The other reason is that all four models share the defining feature of being autoregressive transformers (Radford et al., 2019; Vaswani et al., 2017). Even in this limited analysis, cross-country comparisons consistently favored Great Britain and France opposed to Hungary, which reinforced our initial idea based on published benchmarks, that languages with a larger user base are also more prominently represented in the training data of the models leading to better performance for tasks involving those languages, even though all prompts were given in English.

It must be noted that the reproducibility of the results may be jeopardized by the private companies hosting various versions of the tested models. GPTs are closed source models, OpenAI are both the developers and hosts, while Llama2 and Mixtral are open weights models, which means the trained model is available for anyone to host, but the training data and details of the training process are not open-sourced, prohibiting full reproducibility. This limitation is not of great concern for our research, as we do not aim to pinpoint an exact model and version to be used for the investigated tasks.

A potentially important aspect of the experiment that we are currently ignoring is the time component. The 10th wave of the ESS is a cross-sectional dataset that captures a well-defined timepoint, while the training data for the LLMs span a longer period. The latter is not precisely known but it is at least a decade based on the composition of The Pile dataset (Gao et al., 2020) which is part of the known training data of these models, although the general exponential trend of data created and represented on the internet (Li & Zhang, 2023) suggests that recent data dominate any large enough corpora needed to train such LLMs. We measured the temporal changes in the analyzed distributions looking at five waves of ESS from 6 (dated 2012) to 10 (dated 2022) to assess the potential impact of choosing the latest wave (see Table A7 and A8), and concluded that the values are stable enough in time, and there is no clear reason to prefer any other wave than the 10th or an average of an arbitrary number of waves.

To consider LLMs as viable respondents for survey, some problem-specific data is needed, and providing relevant examples is key – 'Language Models are Few-Shot Learners' as the title of the GPT-3 paper highlights (Brown et al., 2020). Examining the multivariate

associations through the regression coefficients, we found that there is no clear direction of bias for the investigated models. This reinforces the need for application and model specific evaluations. It is currently too early to draw conclusions on the scientific applications in which this methodology proves to be consistently reliable and valid, but there are many avenues for further research, even for this exact testing setup. We plan to expand the analysis of the silicon samples we have by investigating the patterns of answers in more detail, most importantly for open-ended questions. It is clear that models tended to default to certain answers, but it is less obvious whether generating a more faithful answer was dependent on the data structure of ESS (e.g., there is a higher variance in answers for certain personas) or model behavior (e.g., models tend to report higher political interest, because of the characteristics of the training data). To ensure a comprehensive analysis of LLMs, it is of course essential to approach their use in social scientific applications with a critical perspective and to acknowledge the limitations of the linguistic space of the internet in representing social reality. The main language of a country and the representation of that language on the internet most probably influenced the success of the models. Applications of LLMs in social science research should consider the limitations of these models and, if possible, provide solutions by acknowledging and correcting the known biases and caveats.

## References

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies (arXiv:2208.10264). *arXiv.* https://doi.org/10.48550/arXiv.2208.10264

ANES (2021). About Us. *American National Election Studies.* https://electionstudies.org/about-us/

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2022). Out of One, Many: Using Language Models to Simulate Human Samples. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers, pp. 819–862). https://doi.org/10.18653/v1/2022.acl-long.60

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

ESS (2022). About ESS. *European Social Survey.* https://www.europeansocialsurvey.org/about-ess

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S. & Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling (arXiv:2101.00027). *arXiv*. https://doi.org/10.48550/arXiv.2101.00027

Garrido, A. (2009). About some properties of the Kullback–Leibler divergence. *Advanced Modeling and Optimization*, *11*(4), 571–578.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding (arXiv:2009.03300). *arXiv*. https://doi.org/10.48550/arXiv.2009.03300

Li, C. & Zhang, C. (2023). When ChatGPT for Computer Vision Will Come? From 2D to 3D (arXiv:2305.06133). *arXiv*. https://doi.org/10.48550/arXiv.2305.06133

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. (2022). Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do (arXiv:2103.11790). *arXiv*. https://doi.org/10.48550/arXiv.2103.11790

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971). *arXiv*. https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need (arXiv:1706.03762). *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (arXiv:2201.11903). *arXiv*. https://doi.org/10.48550/arXiv.2201.11903

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y. & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models (arXiv:2305.10601). *arXiv*. https://doi.org/10.48550/arXiv.2305.10601

## Funding

The research and work of Zsófia Rakovics was supported by the ÚNKP-23-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development, and Innovation Fund.

The research and work of Zsófia Rakovics was supported by the EKÖP-24 University Excellence Scholarship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

# Appendix

**Table A1** Bootstrap p-values for the pairwise comparisons of KL distributions for the different test setups in the British subsample of ESS.

| | **Llama-2 0shot** | **Mixtral-8x7B 0shot** | **gpt-3.5 0shot** | **gpt-4 0shot** | **Llama-2 5shot** | **Mixtral-8x7B 5shot** | **gpt-3.5 5shot** | **gpt-4 5shot** | **gpt-3.5 5shot rand** | **gpt-4 5shot rand** | **Llama-2 5shot rand** | **Mixtral-8x7B 5shot rand** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Llama-2 0shot | 1 | 0.1365 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mixtral-8x7B 0shot | 0.1365 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| gpt-3.5 0shot | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| gpt-4 0shot | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Llama-2 5shot | 0 | 0 | 0 | 0 | 1 | 0 | 0.0035 | 0 | 0.0095 | 0 | 0 | 0.001 |
| Mixtral-8x7B 5shot | 0 | 0 | 0 | 0 | 0 | 1 | 0.3645 | 0 | 0.403 | 0 | 0 | 0.4525 |
| gpt-3.5 5shot | 0 | 0 | 0 | 0 | 0.0035 | 0.3645 | 1 | 0 | 0.4715 | 0 | 0 | 0.3365 |
| gpt-4 5shot | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| gpt-3.5 5shot rand | 0 | 0 | 0 | 0 | 0.0095 | 0.403 | 0.4715 | 0 | 1 | 0 | 0 | 0.342 |
| gpt-4 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.364 | 0 |
| Llama-2 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.364 | 1 | 0 |
| Mixtral-8x7B 5shot rand | 0 | 0 | 0 | 0 | 0.001 | 0.4525 | 0.3365 | 0 | 0.342 | 0 | 0 | 1 |

**Table A2** Bootstrap p-values for the pairwise comparisons of KL distributions for the different test setups in the French subsample of ESS.

| | **Llama-2 0shot** | **Mixtral-8x7B 0shot** | **gpt-3.5 0shot** | **gpt-4 0shot** | **Llama-2 5shot** | **Mixtral-8x7B 5shot** | **gpt-3.5 5shot** | **gpt-4 5shot** | **gpt-3.5 5shot rand** | **gpt-4 5shot rand** | **Llama-2 5shot rand** | **Mixtral-8x7B 5shot rand** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Llama-2 0shot | 1 | 0.0525 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mixtral-8x7B 0shot | 0.0525 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| gpt-3.5 0shot | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| gpt-4 0shot | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Llama-2 5shot | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mixtral-8x7B 5shot | 0 | 0 | 0 | 0 | 0 | 1 | 0.0755 | 0 | 0 | 0 | 0 | 0 |
| gpt-3.5 5shot | 0 | 0 | 0 | 0 | 0 | 0.0755 | 1 | 0 | 0 | 0 | 0 | 0 |
| gpt-4 5shot | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| gpt-3.5 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.0025 | 0 |
| gpt-4 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Llama-2 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0025 | 0 | 1 | 0.0025 |
| Mixtral-8x7B 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0025 | 1 |

**Table A3** Bootstrap p-values for the pairwise comparisons of KL distributions for the different test setups in the Hungarian subsample of ESS.

| | **Llama-2 0shot** | **Mixtral-8x7B 0shot** | **gpt-3.5 0shot** | **gpt-4 0shot** | **Llama-2 5shot** | **Mixtral-8x7B 5shot** | **gpt-3.5 5shot** | **gpt-4 5shot** | **gpt-3.5 5shot rand** | **gpt-4 5shot rand** | **Llama-2 5shot rand** | **Mixtral-8x7B 5shot rand** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Llama-2 0shot | 1 | 0.1875 | 0 | 0.044 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mixtral-8x7B 0shot | 0.1875 | 1 | 0.007 | 0.1415 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| gpt-3.5 0shot | 0 | 0.007 | 1 | 0.1645 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| gpt-4 0shot | 0.044 | 0.1415 | 0.1645 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Llama-2 5shot | 0 | 0 | 0 | 0 | 1 | 0 | 0.0005 | 0 | 0 | 0 | 0 | 0 |
| Mixtral-8x7B 5shot | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| gpt-3.5 5shot | 0 | 0 | 0 | 0 | 0.0005 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| gpt-4 5shot | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| gpt-3.5 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.0425 | 0 | 0 |
| gpt-4 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0425 | 1 | 0 | 0 |
| Llama-2 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.477 |
| Mixtral-8x7B 5shot rand | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.477 | 1 |

**Table A4** Standardized regression coefficients estimated from the ESS Great Britain subsample and the generated samples of the models (GPT-3.5-turbo, GPT-4-turbo, Llama-2-70b, Mixtral-8x7B) with different prompts (0-shot, 5-shot with random examples, 5-shot with closest examples from the data for the given characteristics) for political interest of the participants with gender, age, education, and political preferences on the left-right scale as explanatory variables.

| Setup | Gender | SE Gender | Age | SE Age | Education | SE Education | LR-scale | SE LR-scale |
|---|---|---|---|---|---|---|---|---|
| ESS GB | 0.088 | 0.027 | −0.197 | 0.030 | −0.290 | 0.029 | 0.073 | 0.030 |
| Llama-2 0shot | 0.036 | 0.029 | −0.132 | 0.032 | −0.106 | 0.032 | 0.122 | 0.049 |
| Mixtral-8x7B 0shot | 0.042 | 0.029 | −0.031 | 0.032 | −0.003 | 0.031 | 0.093 | 0.037 |
| gpt-3.5 0shot | 0.002 | 0.029 | −0.086 | 0.028 | 0.006 | 0.031 | 0.111 | 0.061 |
| gpt-4 0shot | 0.070 | 0.030 | −0.080 | 0.029 | −0.079 | 0.032 | 0.110 | 0.053 |
| Llama-2 5shot | 0.091 | 0.026 | −0.306 | 0.028 | −0.315 | 0.028 | 0.192 | 0.033 |
| Mixtral-8x7B 5shot | 0.089 | 0.027 | −0.207 | 0.029 | −0.370 | 0.029 | 0.105 | 0.036 |
| gpt-3.5 5shot | 0.029 | 0.027 | −0.150 | 0.027 | −0.403 | 0.027 | 0.100 | 0.029 |
| gpt-4 5shot | 0.098 | 0.027 | −0.190 | 0.028 | −0.368 | 0.028 | 0.073 | 0.035 |
| gpt-3.5 5shot rand | 0.041 | 0.027 | −0.052 | 0.030 | −0.344 | 0.029 | 0.168 | 0.033 |
| gpt-4 5shot rand | 0.075 | 0.028 | −0.044 | 0.031 | −0.350 | 0.031 | 0.101 | 0.044 |
| Llama-2 5shot rand | 0.031 | 0.028 | −0.013 | 0.028 | −0.171 | 0.031 | 0.153 | 0.040 |
| Mixtral-8x7B 5shot rand | 0.083 | 0.026 | −0.035 | 0.030 | −0.373 | 0.029 | 0.142 | 0.044 |

**Table A5** Standardized regression coefficients estimated from the ESS France subsample and the generated samples of the models (GPT-3.5-turbo, GPT-4-turbo, Llama-2-70b, Mixtral-8x7B) with different prompts (0-shot, 5-shot with random examples, 5-shot with closest examples from the data for the given characteristics) for political interest of the participants with gender, age, education, and political preferences on the left-right scale as explanatory variables.

| Setup | Gender | SE Gender | Age | SE Age | Education | SE Education | LR-scale | SE LR-scale |
|---|---|---|---|---|---|---|---|---|
| ESS FR | 0.133 | 0.020 | −0.205 | 0.019 | −0.347 | 0.021 | 0.020 | 0.023 |
| Llama-2 0shot | −0.043 | 0.022 | −0.010 | 0.024 | −0.132 | 0.025 | 0.073 | 0.037 |
| Mixtral-8x7B 0shot | 0.016 | 0.023 | 0.014 | 0.024 | −0.036 | 0.022 | 0.015 | 0.027 |
| gpt-3.5 0shot | 0.037 | 0.022 | −0.158 | 0.022 | 0.003 | 0.024 | 0.082 | 0.037 |
| gpt-4 0shot | 0.015 | 0.022 | −0.144 | 0.021 | 0.003 | 0.025 | 0.036 | 0.032 |
| Llama-2 5shot | 0.122 | 0.020 | −0.247 | 0.019 | −0.376 | 0.019 | 0.079 | 0.026 |
| Mixtral-8x7B 5shot | 0.127 | 0.020 | −0.173 | 0.019 | −0.434 | 0.020 | 0.072 | 0.027 |
| gpt-3.5 5shot | 0.066 | 0.020 | −0.167 | 0.020 | −0.424 | 0.019 | 0.035 | 0.022 |
| gpt-4 5shot | 0.150 | 0.019 | −0.223 | 0.019 | −0.440 | 0.019 | 0.054 | 0.026 |
| gpt-3.5 5shot rand | 0.007 | 0.020 | 0.038 | 0.023 | −0.346 | 0.021 | 0.152 | 0.024 |
| gpt-4 5shot rand | 0.087 | 0.021 | −0.043 | 0.022 | −0.370 | 0.021 | 0.089 | 0.030 |
| Llama-2 5shot rand | −0.009 | 0.022 | 0.008 | 0.023 | −0.229 | 0.021 | 0.179 | 0.030 |
| Mixtral-8x7B 5shot rand | 0.090 | 0.019 | −0.012 | 0.021 | −0.418 | 0.018 | 0.130 | 0.031 |

**Table A6** Standardized regression coefficients estimated from the ESS Hungary subsample and the generated samples of the models (GPT-3.5-turbo, GPT-4-turbo, Llama-2-70b, Mixtral-8x7B) with different prompts (0-shot, 5-shot with random examples, 5-shot with closest examples from the data for the given characteristics) for political interest of the participants with gender, age, education, and political preferences on the left-right scale as explanatory variables

| Setup | Gender | SE Gender | Age | SE Age | Education | SE Education | LR-scale | SE LR-scale |
|---|---|---|---|---|---|---|---|---|
| ESS HU | 0.134 | 0.021 | −0.261 | 0.022 | −0.218 | 0.022 | −0.094 | 0.023 |
| gpt-3.5 0shot | 0.006 | 0.023 | −0.084 | 0.021 | −0.080 | 0.022 | −0.157 | 0.039 |
| gpt-4 0shot | −0.009 | 0.023 | −0.061 | 0.023 | −0.101 | 0.023 | −0.174 | 0.039 |
| Llama-2 0shot | −0.022 | 0.023 | −0.056 | 0.024 | −0.031 | 0.024 | −0.124 | 0.037 |
| Mixtral-8x7B 0shot | 0.051 | 0.024 | 0.014 | 0.024 | −0.141 | 0.022 | −0.014 | 0.026 |
| gpt-3.5 5shot | 0.089 | 0.021 | −0.187 | 0.024 | −0.346 | 0.021 | −0.039 | 0.023 |
| gpt-4 5shot | 0.157 | 0.020 | −0.288 | 0.021 | −0.332 | 0.020 | −0.065 | 0.027 |
| Llama-2 5shot | 0.085 | 0.021 | −0.280 | 0.023 | −0.266 | 0.021 | −0.016 | 0.028 |
| Mixtral-8x7B 5shot | 0.156 | 0.021 | −0.191 | 0.023 | −0.377 | 0.020 | −0.011 | 0.024 |
| gpt-3.5 5shot rand | 0.030 | 0.022 | −0.026 | 0.024 | −0.290 | 0.022 | 0.070 | 0.027 |
| gpt-4 5shot rand | 0.116 | 0.021 | −0.067 | 0.023 | −0.305 | 0.019 | −0.054 | 0.030 |
| Llama-2 5shot rand | 0.000 | 0.023 | −0.014 | 0.023 | −0.188 | 0.022 | 0.049 | 0.033 |
| Mixtral-8x7B 5shot rand | 0.103 | 0.021 | −0.015 | 0.022 | −0.365 | 0.020 | 0.030 | 0.028 |

**Table A7** Descriptive statistics of the distribution of the political interest variable from ESS waves 6 to 10

| **Country** | **Great Britain** | | **France** | | **Hungary** | |
|---|---|---|---|---|---|---|
| Answer category | ESS10 | Mean (SD) of waves 6 to 10 | ESS10 | Mean (SD) of waves 6 to 10 | ESS10 | Mean (SD) of waves 6 to 10 |
| very interested | 20.1% | 15.9% (2.6%) | 15.1% | 15.8% (1.2%) | 3.6% | 4.1% (0.6%) |
| quite interested | 43.8% | 41.4% (2.2%) | 24.2% | 28.7% (4.3%) | 21.5% | 23.1% (1.9%) |
| hardly interested | 22.0% | 25.2% (2.0%) | 40.5% | 36.9% (2.8%) | 45.0% | 40.0% (3.5%) |
| not at all interested | 14.0% | 17.5% (3.1%) | 20.2% | 18.6% (1.1%) | 29.8% | 32.8% (2.4%) |

**Table A8** Point estimates of regression coefficients from the 10th compared to the mean (SD) coefficients from waves 6 to 10.

| **Country** | **Great Britain** | | **France** | | **Hungary** | |
|---|---|---|---|---|---|---|
| Variable | ESS10 estimate | Mean (SD) of waves 6 to 10 | ESS10 estimate | Mean (SD) of waves 6 to 10 | ESS10 estimate | Mean (SD) of waves 6 to 10 |
| Gender | 0.168 | 0.215 (0.059) | 0.258 | 0.266 (0.023) | 0.224 | 0.245 (0.078) |
| Age | −0.010 | −0.009 (0.001) | −0.011 | −0.009 (0.001) | −0.011 | −0.010 (0.001) |
| Education | −0.207 | −0.221 (0.027) | −0.293 | −0.259 (0.019) | −0.230 | −0.209 (0.030) |
| LR-scale | 0.033 | −0.004 (0.023) | 0.009 | 0.001 (0.009) | −0.034 | −0.044 (0.011) |