Creating Artificial Survey Panels Is Still Difficult

Gabriela Kadlecová[1,2,3][0000−0002−4780−0633], Petra Vidnerová[1,2][0000−0003−3879−3459], Roman Neruda[1,2][0000−0003−2364−5357], and Josef Šlerka[1][0000−0003−3564−1767]

[1] Charles University, Faculty of Arts
[2] The Czech Academy of Sciences, Institute of Computer Science
[3] Charles University, Faculty of Mathematics and Physics

Abstract. Large language models have shown the ability to emulate human behavior and opinions, suggesting their potential use in creating artificial survey panels. This paper revisits a previous work on German 2017 parliamentary elections and expands this approach to the Czech parliamentary elections of 2021, which involves a more fragmented political landscape and potentially less coverage in English sources. We evaluate two recent models and examine predictions on both subgroup and individual levels. Our findings suggest that although the models improved across a wide range of language tasks, creating accurate artificial survey panels is still difficult.

Keywords: Artificial Survey Panel · Large Language Model · Elections

1 Introduction

Over the past year, *large language models (LLM)* have become prevalent in various applications from high-quality chatbots, over reasoning to translation or text generation tasks [10, 15]. The models were also shown to model human behavior, morality, and opinions [1, 3, 9], thus the further natural application is artificial survey panels. Previous work by [12] has shown that the generated survey panels are prone to biases and inaccurate individual-level predictions. Ever since, many models with better performance across various tasks have emerged [6, 7, 13, 15].

In this paper, we revisit the German study by von der Heyde et al. [12], and examine whether predicting the *distribution* of votes and using a more powerful model leads to a better survey panel votes estimation. Then, we extend the work to 2021 parliamentary elections data in the Czech Republic – a country with multiple political parties less covered in English internet sources. We examine model predictions on subgroup and individual levels by the Command r+ and Mixtral models. Based on the results, we discuss directions for future research on artificial survey panels.

2 Related Work

The large language models are a recent breakthrough in natural language processing applications. Since the early models [8], different closed-source and open-source (with published weights, i.e. available for local use) models have emerged. Examples of proprietary closed-source models include GPT-3, GPT-4 [5, 15], or Claude 3.5 Sonnet [2]. Among models with open weights, there are Gemma [10], Command r+ [7], Mixtral [13], or LLaMA [20].

GPT-3.5 was shown to accurately model human morality by Dillion et al. [9]. They argue that the models can simulate some human behavior and choices, and could be useful, especially in the early stages of the survey development [9]. Another work examined that the models faithfully simulate human behavior in classic economic, psycholinguistic, and social psychology experiments [1].

Argyle et al. [3] used GPT-3 to simulate artificial voter conditioned on sociodemographic profiles from several US surveys. Their system was able to estimate voting preferences in US elections with high fidelity. A similar study of von der Heyde et al. [12] created an artificial survey panel based on the German Longitudinal Election Study (GLES).

Many shortcomings of LLM have been identified, such as the lack of interpretability and difficulty to compare with traditional statistical approaches. Authors of [4] compare statistical models and artificial panels for 2016–2020 American National Election Study (ANES) data, and observe lower variance, dependency on prompt wording, and inconsistent outputs in time. Recent study [22] revisits the results of [3] and identifies shortcuts in data – two input features that predict the output with high reliability. By removing the shortcuts, the accuracy of the panel dramatically decreases.

The above mentioned problems are addressed in the work of Kin and Lee that fine-tune the Alpaca7b LLM with cross-sectional surveys to incorporate the meaning of survey questions, individual beliefs, and temporal contexts [14]. Pham and Cunningham [16] showed that narrative prompting can help the GPT models predict future events after the data cutoff.

3 Methods

In this paper, we use large language models to create an artificial survey panel,[4] specifically, we use the model to predict what parties the respondent voted for. Our work is similar to the artificial survey panel created by von der Heyde et al. [12] and based on the German Longitudinal Election Study (GLES). Compared to the study, we used only a subset of the data, but we use more recent large language models, and we output the *distribution* for potential parties the respondent might have voted for. We also extended the work to Czech election data from the Society of Distrust dataset. In the following subsections, we describe our setup.

Just as the original paper [12], we use the post-election cross-section GLES 2017 data [11]. We used the same respondents and made a slight change to the individual characteristics – we removed the partisanship, as it serves as a shortcut for the output. Additionally, we added the ‘fear of refugees’ feature, as it improved the results in initial experiments.

[4] Code is publicly available at `https://github.com/gabikadlecova/panelart`

For our experiments on the Czech parliamentary elections 2021, we used open data from the *Society of distrust*[5] project [17]. The aim of the project was to analyze the prevalence of conspiracy and disinformation narratives in Czech society via a questionnaire. Along with basic socio-economical data, it included questions on trust in institutions, what party the respondent voted for in the parliamentary elections 2021, and other questions related to belief in conspiracies. The data collection was done in 2023, meaning the election responses are retrospective.

The study contains 3880 respondent entries with 170 covariates per respondent. Out of these, we selected 13 respondent data columns and 1 target column (2021 election vote). We summarize the data in Table 1.

Table 1. Society of Distrust – columns selected as input data .

column	description
gender	Male/female [6]
age	people aged 18-65
education	high-school and above
kraj	region – all 13 regions + Prague
okres	district – subdivision of regions
town size	5 town size categories
employment	multiple variants of employment[7]
income	8 income categories
zivotni_uroven	self-proclaimed quality of life
zajem_politika	interest in politics
eu	satisfaction with EU membership
nato	satisfaction with NATO membership
covid_vacc	whether the respondent received the COVID-19 vaccine

For our experiments, we used two large language models, the Command r+ [7] through proprietary API, and the local Mixtral model [13]. Based on the LMSYS Chatbot Arena Leaderboard [6], both models outperform older versions of GPT-3.5, the Command r+ also outperforms older versions of GPT-4.

As the input, we used prompts similar to those in the original study [12] with the following modification. Instead of predicting 1 party only, we let the model output probabilities of multiple parties. Then, we predict the vote by sampling from the party distribution. To do so, we add instructions at the beginning of the prompt:

`In place of [INSERT], fill in (in German) whether the respondent voted and if yes, then for what party. If unsure, list the probable parties with probabilities (always output whether the respondent voted and for what party). List as many parties as necessary. The probabilities should sum up to 1.`

[5] The data is publicly available after registration.

[6] Third gender is not officially recognized in Czechia.

[7] Including retirement, childcare, and students.

Optionally answer "andere Partei" if the voter voted for a small unpopular party. The output format is: [gewählt, proba a], [nicht gewählt, proba b]; [PARTY1, proba 1], [PARTY2, proba 2],...*

Based on the output quality, we added more instructions on the correct format (e.g. where the semicolon should be). The respondent part of the prompt is the same as in the original study except for the change in partisanship and addition of the 'fear of refugees' column.

For the Czech elections, we used the same prompt design, and replaced the respondent part according to the Society of Distrust data. An example prompt translated from Czech:

I am a woman, 49 years old, with a secondary education diploma. I live in the Ústí nad Labem region, in a town with a population of 5,001 - 20,000. In terms of employment, I am a homemaker, and our household income is less than 15,000 CZK. I am rather dissatisfied that the Czech Republic is a member state of NATO. I have neither a good nor a bad standard of living. I am somewhat interested in politics. I am not vaccinated against COVID-19. In the 2021 parliamentary elections, I [INSERT].

Example of the model output (the probabilities needed to be normalized):

* [volil, 0.6], [nevolil, 0.4]; [ANO, 0.25], [SPOLU, 0.15], [Piráti a STAN, 0.1], [Piráti, 0.05], [STAN, 0.05], [SPD, 0.15], [KSČM, 0.05], [ČSSD, 0.05], [jiná strana, 0.1]

4 Experiments

We first evaluated the Command r+ model on the German parliamentary elections 2017 data to compare them to the GPT-3 model from [12]. Afterward, we evaluated two models on the Czech parliamentary elections 2021.

4.1 German Parliamentary Elections 2017

The results of Command r+ on GLES 2017 data are depicted in Figure 1. Compared to the study by von der Heyde et al. [12], the model overestimates 'CDU/CSU', 'AfD', 'a different party' (that their model underestimated), but underestimates the fraction of non-voters and 'Grüne' (that their model overestimated). The other 3 predictions are similar. This means that no significant improvement occurred with a newer, more capable model, but with a possible shortcut removed from data.

An interesting observation about the possible bias in data caused by their availability can be demonstrated by comparing the internet presence of parties. Figure 2, taken from Serrano et al. [19] shows statistics of tweets aggregated by the AfD party. In Figure 3 we visualize the distribution of tweets by party affiliation according to [18] It is unknown what data large language models were trained on. However, we can see that FDP has a lower share of available data in both cases, while AfD and Die Linke have a larger share. Also, the scraped tweets in Figure 2 show a lower share of tweets about the Green party. It could

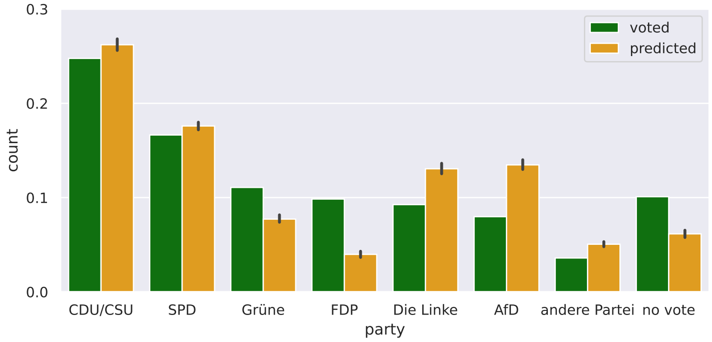

Fig. 1. 2017 vote prediction via Command r+ – all respondents from GLES. Last 2 columns – other party, non-voters.

be possible that the availability of data affected the model predictions, as it closely resembles the predicted distribution in Figure 1.

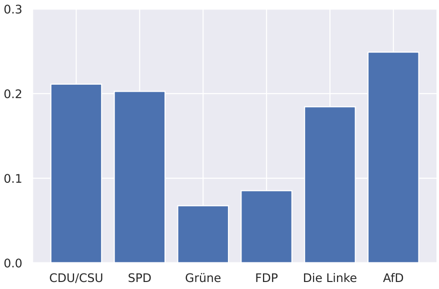

Fig. 2. Distribution of randomly scraped tweets in 2017 by Serrano et al. [19].

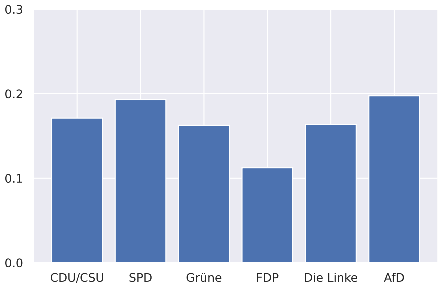

Fig. 3. Distribution of selected 2021 elections tweets of politicians by [18]

4.2 Czech Parliamentary Elections 2021

There are multiple smaller parties in the Czech political system, and moreover, the 2021 parliamentary elections were unique in that 2 large coalitions were formed – SPOLU from 3 parties (ODS, TOP 09, and KDU-ČSL) and a second one from the Czech Pirate Party and STAN. This is more complex compared to Germany, where only the AfD is a recent addition to the system.

First, we used the Command r+ model to predict the vote distribution for all respondents in the Society of Distrust data. Figure 4 shows the predicted fraction of all votes (with 95% confidence intervals computed from 10 party selections),

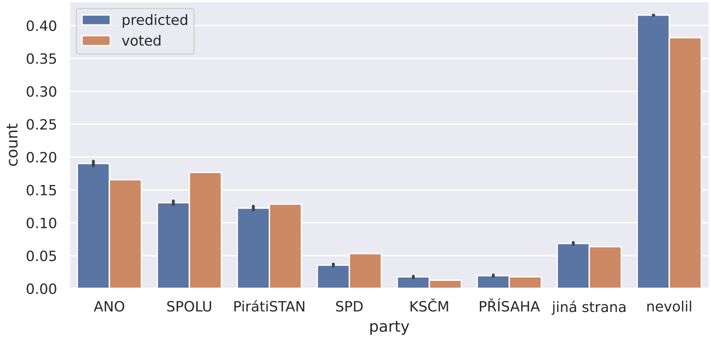

Fig. 4. 2021 Vote prediction via Command r+ – all respondents from the Society of Distrust. Last 2 columns – other party, non-voters.

and the fraction of reported votes in the study. We see that the model quite accurately predicts votes for smaller parties, yet it does not order the top 3 well. A possible explanation could be that before the elections, survey panels favored ANO, and the coalition of Pirates and STAN had a drop in support just before the elections [21].

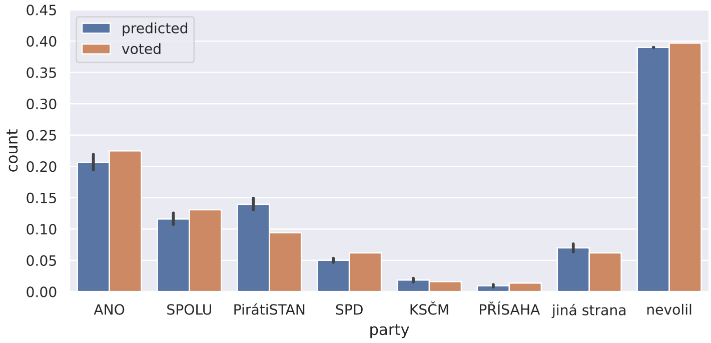

Fig. 5. 2021 vote prediction via Command r+ – respondents from the Moravskoslezsky region. Last 2 columns – other party, non-voters.

Next, we looked at subgroups of respondents. The first group are respondents from the Moravskoslezsky region – a region, where the support for the two coalitions was lower, and the support for ANO higher (Figure 5, 436 respondents). The second group were people aged 18-24 years (Figure 6, 397 respondents). We see that the predictions are similar to the results on the full data, and do not reflect the real votes well. This indicates that either the model is not able

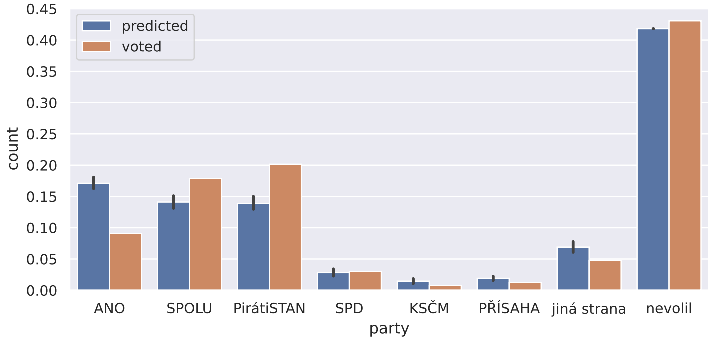

Fig. 6. 2021 vote prediction via Command r+ – respondents aged 18-24 years. Last 2 columns – other party, non-voters.

to capture the nuances of subgroups, or that the input data misses information that would help to disaggregate the respondents.

We also examined how accurately the model predicts individual respondent votes. The model estimated the correct party/non-vote only in 25% of the cases, and out of voters, the correct party was estimated only in 14% of cases. As so, the model captures mostly global voting trends rather than fine-grained predictions.

For the last batch of experiments performed with the Mixtral model, we randomly selected 200 respondents due to the long prediction time of Mixtral. We used two variants of Mixtral – small (`Mixtral-8x7B`), and large (`Mixtral-8x22B`). Figure 7 shows the results of the smaller model. We see that the predicted results are poor – SPD has a much higher, SPOLU much lower support, the non-voters are grossly underestimated. For the larger model (Figure 8), the results are slightly better but non-voters are still underestimated, and 'other party' is overestimated. An explanation could be that a low amount of Czech sources was present in the training data, or that this particular task could be hard for this model.

5 Conclusion

Inspired by recent work by von der Heyde et al. [12] we analyzed the possibility of creating an artificial survey panel for Czech election data. Our approach is novel in using more recent models, and in predicting vote distribution rather than 1 party vote.

We found that Command r+ [7], a more recent model that outperforms GPT-3 [5], produces inaccurate predictions that seem to follow the distribution of online posts about parties. On the Czech parliamentary election data, Command r+ was able to accurately estimate the votes for smaller parties and non-voters, yet it failed to predict the ranking of top 3 parties/coalitions. When doing predictions on subgroups, the model exhibited poor results, indicating a

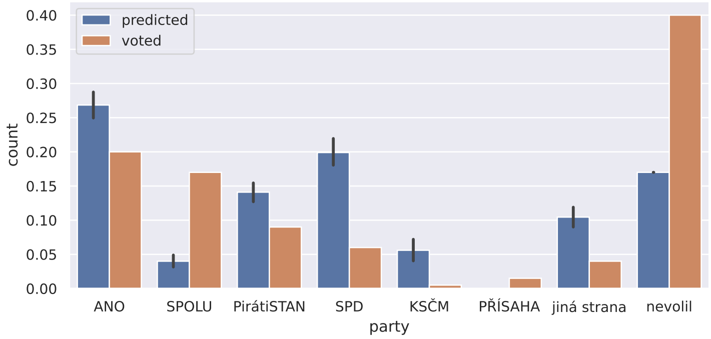

Fig. 7. 2021 vote prediction via Mixtral (small), 200 respondents. Last 2 columns – other party, non-voters.

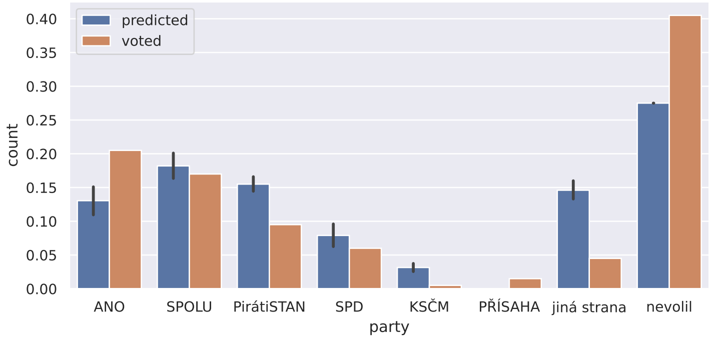

Fig. 8. 2021 vote prediction via Mixtral (large), 200 respondents. Last 2 columns – other party, non-voters.

need for more detailed training data, or an overall inability to predict in finer granularities. Finally, we showed that Mixtral [13], another recent model better than GPT-3, was unable to match any part of the reference survey panel.

Overall, we showed that although recent models outperform first large language models across a range of tasks, the task of predicting elections in a small country with several parties is hard. Future work should focus on analyzing the source of the bias, examine sensitivity to small variations in the prompt, and include other recent models in the evaluation (namely the GPT-4[15]).

To improve the accuracy, models should exhibit a better performance on subgroup and individual levels. Since other works showed that the models can model human behavior well, data that describes human personalities could be a key to achieving more accurate predictions. The models could also be fine-tuned to publicly available Czech news data, as it is uncertain how present they were in model training input datasets.

The unavailability of the data on which large language models were trained is the main limitation of our approach. Additionally, the version of closed-source models can change during experimentation, leading to inconsistent results.

Acknowledgments. This work was supported by the European Regional Development Fund project "Beyond Security: Role of Conflict in Resilience-Building" (reg. no.: CZ.02.01.01/00/22_008/0004595). This research was partially supported by SVV project number 260 698.

Disclosure of Interests. The authors have no competing interests.

References

1. Aher, G., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org (2023)
2. Anthropic: Introducing claude 3.5 sonnet: A new standard in ai performance. `https://www.anthropic.com/news/claude-3-5-sonnet` (2024), [Accessed 19-07-2024]
3. Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: Using language models to simulate human samples. Political Analysis **31**(3), 337–351 (2023), https://doi.org/10.1017/pan.2023.2
4. Bisbee, J., Clinton, J.D., Dorff, C., Kenkel, B., Larson, J.M.: Synthetic replacements for human survey data? the perils of large language models. Political Analysis p. 1–16 (2024), https://doi.org/10.1017/pan.2024.5
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D.e.a.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, Curran Associates, Inc. (2020), URL `https://proceedings.neurips.cc/paper\_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`
6. Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J.E., Stoica, I.: Chatbot arena: An open platform for evaluating llms by human preference (2024), URL `https://arxiv.org/abs/2403.04132`
7. Cohere: Command R+ — docs.cohere.com. `https://docs.cohere.com/docs/command-r-plus` (2024), [Accessed 18-07-2024]
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://doi.org/10.18653/v1/N19-1423, URL `https://aclanthology.org/N19-1423`

9. Dillion, D., Tandon, N., Gu, Y., Gray, K.: Can ai language models replace human participants? Trends in Cognitive Sciences **27**(7), 597–600 (2023), ISSN 1364-6613, https://doi.org/https://doi.org/10.1016/j.tics.2023.04.008, URL `https://www.sciencedirect.com/science/article/pii/S1364661323000980`
10. Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., et al., S.B.: Gemma: Open models based on gemini research and technology (2024), URL `https://arxiv.org/abs/2403.08295`
11. GLES: Pre- and post-election cross section (cumulation) (gles 2017). GESIS Data Archive, Cologne. ZA6802 Data file Version 3.0.1, https://doi.org/10.4232/1.13236 (2019), https://doi.org/10.4232/1.13236
12. von der Heyde, L., Haensch, A.C., Wenz, A.: Assessing bias in llm-generated synthetic datasets: The case of german voter behavior (Dec 2023), https://doi.org/10.31235/osf.io/97r8s, URL `osf.io/preprints/socarxiv/97r8s`
13. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., et al., B.S.: Mixtral of experts (2024), URL `https://arxiv.org/abs/2401.04088`
14. Kim, J., Lee, B.: Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction (2024), URL `https://arxiv.org/abs/2305.09620`
15. OpenAI, Achiam, J., Adler, S., Agarwal, S., et al., L.A.: Gpt-4 technical report (2024), URL `https://arxiv.org/abs/2303.08774`
16. Pham, V., Cunningham, S.: Can base chatgpt be used for forecasting without additional optimization? (2024), URL `https://arxiv.org/abs/2404.07396`
17. Pilnáček, M., Tabery, P., Šlerka, J., Heřmanová, M., Šimoník, P., Žáčková, L.: Society of Distrust 2023 (2024), https://doi.org/10.14473/CSDA/G57AR8, URL `https://doi.org/10.14473/CSDA/G57AR8`
18. Schmidt, T., Fehle, J., Weissenbacher, M., Richter, J., Gottschalk, P., Wolff, C.: Sentiment analysis on twitter for the major german parties during the 2021 german federal election. In: KONVENS, pp. 74–87 (2022), URL `https://aclanthology.org/2022.konvens-1.9`
19. Serrano, J.C.M., Shahrezaye, M., Papakyriakopoulos, O., Hegelich, S.: The rise of germany's afd: A social media analysis. In: Proceedings of the 10th International Conference on Social Media and Society, p. 214–223, SMSociety '19, Association for Computing Machinery, New York, NY, USA (2019), ISBN 9781450366519, https://doi.org/10.1145/3328529.3328562, URL `https://doi.org/10.1145/3328529.3328562`
20. 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.: Llama: Open and efficient foundation language models (2023), URL `https://arxiv.org/abs/2302.13971`
21. Volební kalkulačka Team: Mandáty.cz - aktualizovaný model pro volby do Sněmovny — 2021.mandaty.cz. `https://2021.mandaty.cz` (2021), [Accessed 19-07-2024, available in Czech]
22. Yang, K., Li, H., Wen, H., Peng, T.Q., Tang, J., Liu, H.: Are large language models (llms) good social predictors? (2024), URL `https://arxiv.org/abs/2402.12620`