Check for updates

Original Manuscript

Social Science Computer Review
2025, Vol. 0(0) 1–23
© The Author(s) 2025

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393251337014
journals.sagepub.com/home/ssc

Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice

Leah von der Heyde[1,2], Anna-Carolina Haensch[1,3], and Alexander Wenz[4]

Abstract

"Synthetic samples" generated by large language models (LLMs) have been argued to complement or replace traditional surveys, assuming their training data is grounded in human-generated data that potentially reflects attitudes and behaviors prevalent in the population. Initial US-based studies that have prompted LLMs to mimic survey respondents found that the responses match survey data. However, the relationship between the respective target population and LLM training data might affect the generalizability of such findings. In this paper, we critically evaluate the use of LLMs for public opinion research in a different context, by investigating whether LLMs can estimate vote choice in Germany. We generate a synthetic sample matching the 2017 German Longitudinal Election Study respondents and ask the LLM GPT-3.5 to predict each respondent's vote choice. Comparing these predictions to the survey-based estimates on the aggregate and subgroup levels, we find that GPT-3.5 exhibits a bias towards the Green and Left parties. While the LLM predictions capture the tendencies of "typical" voters, they miss more complex factors of vote choice. By examining the LLM-based prediction of voting behavior in a non-English speaking context, our study contributes to research on the extent to which LLMs can be leveraged for studying public opinion. The findings point to disparities in opinion representation in LLMs and underscore the limitations in applying them for public opinion estimation.

Keywords

LLM, artificial intelligence, survey data, synthetic samples, voting behavior, public opinion

[1]LMU Munich, Germany
[2]Munich Center for Machine Learning, Germany
[3]University of Maryland, USA
[4]University of Mannheim, Germany

Corresponding Author:
Leah von der Heyde, Department of Statistics, Social Data Science & AI Lab (Kreuter), LMU Munich, Ludwigstr. 33, München 80359, Germany.
Email: l.heyde@lmu.de

Introduction

The recent development and large-scale proliferation of large language models (LLMs), such as OpenAI's GPT (OpenAI et al., 2023) or Meta's Llama (Touvron et al., 2023), have spurred discussions about the extent to which these language models can be used for research in the social and behavioral sciences. Researchers have started exploring various applications to facilitate the collection and analysis of survey data. Examples include the use of LLMs for questionnaire design and scale development (Götz et al., 2023; Hernandez & Nie, 2022; Konstantis et al., 2023; Laverghetta & Licato, 2023; Lee et al., 2023), conducting interviews (Chopra & Haaland, 2023; Villalba et al., 2023), coding open-ended survey responses (Mellon et al., 2024; Rytting et al., 2023), imputing missing data and detecting statistical outliers (Jaimovitch-López et al., 2023; Kim & Lee, 2023), detecting non-human respondents in online surveys (Lebrun et al., 2023), and data visualization and interpretation (Liew & Mueller, 2022; Sultanum & Srinivasan, 2023).

Beyond augmenting survey data collection and analysis, research has also started to examine to what extent LLMs can be used for making valid inferences about a population (e.g., Argyle et al., 2023). LLMs are trained on large amounts of internet text data, such as selected book collections, Wikipedia, and social media data, which are assumed to reflect attitudes and behaviors prevalent in the population. Their text output to a request represents a conditional probability based on the training data and the specific contextual information provided in the request. Thus, some researchers have proposed that *synthetic samples* generated by LLMs might serve as a novel, fast, and cost-efficient method of collecting data about public opinion—provided they yield similar estimates and correlates as existing data collection methods. Such samples have been created by sequentially feeding individual socio-demographic, socio-economic, and/or attitudinal information of specific persons to an LLM and asking it to respond to survey questions from the respective person's perspective.

There has been an increasing number of academic studies and non-academic applications using LLMs for population inference. In light of initial studies showing that synthetic samples match survey data (e.g., Argyle et al., 2023; Chu et al., 2023), a surge of startups have started offering "solutions" based on LLM-synthetic samples (e.g., Aaru, n.d.; Delve AI, n.d.; Synthetic Users, n.d.). However, more recent research challenges the initial scientific findings when comparing LLM-based data to that of surveys (e.g., Bisbee et al., 2024; Dominguez-Olmedo et al., 2023; Santurkar et al., 2023). Nevertheless, some of these organizations continue to capitalize on the challenge of predicting public opinion, particularly voting behavior. For example, an AI startup has entered the polling race, trying (and failing) to predict the 2024 US election using synthetic samples (Chua, 2024; Mendoza, 2024). This dynamic could have far-reaching consequences for the polling industry. In light of this, it is important to systematically investigate the biases in LLM-based synthetic public opinion data.

However, most existing research using *individual-level* synthetic samples, whether positive or negative in its findings, has focused on the United States. It is thus unclear how LLMs perform in estimating individual-level public opinion in other political, cultural, and linguistic contexts. We argue that the suitability of LLMs for estimating public opinion outside the US population is even more questionable than within the US, as their effectiveness may depend on various contextual factors associated with the target population. These factors include (1) the prevalence of native-language training data, (2) a country's political and societal structure, which has a complex relationship with public opinion that can vary across countries and might not be equally reflected in the training data, as well as (3) structural differences between the target population and the population reflected in the training data. However, details about LLM "inputs"—such as their architectures and training datasets—are largely inaccessible to the scientific community, especially for proprietary models like those from OpenAI. Studies such as Ball et al. (2024), which do

focus on examining internal model components, have a very limited scope, highlighting the challenges of generalizing findings beyond the specific examples studied. Therefore, at least as a first step, researchers investigate differences and biases in LLM outputs as a proxy. For example, using logic tasks, McCoy et al. (2023) demonstrate that LLM output is skewed towards tasks that are known to be more commonly mentioned in internet text, suggesting biases in LLM training data can indeed be proxied through its output. In our case, this implies analyzing the LLM-generated public opinion data. Polling voting behavior is one relevant and much-researched example of public opinion estimation. It is also an example that is heavily dependent on the national social and political context. For example, the dynamics of vote choice are markedly different in a multi-party system, such as Germany's, than in the US two-party system. At the same time, due to its linguistic and socio-demographic presence online and its socio-political structure, Germany presents a reasonable middle ground for the examination of LLM-based public opinion estimation, the results of which can be telling for societies represented in LLM training data even less.

In this paper, we examine to what extent LLMs can estimate public opinion in Germany by addressing the following research questions:

RQ1. Do LLM-based samples provide similar estimates of voting behavior as national election studies?
RQ2. How do LLMs' estimates of voting behavior deviate from national election studies for different subgroups of the population?

Following the approach employed by Argyle et al. (2023), we create a synthetic sample of eligible voters based on data from the German Longitudinal Election Study (GLES). These personas include individual-level information on variables that in the literature have been found to be important predictors of voting behavior—demographics, party affiliations, and views on politically salient issues, such as immigration. Based on this information, we prompt the LLM GPT-3.5 in German to predict the voting behavior of each individual. From the LLM responses, we extract the predicted vote choices for each persona and compare them to the voting behavior reported by respondents in the GLES data. Thus, our primary goal in this paper is not to assess whether LLMs can predict actual election outcomes, but whether they can infer individual voting behavior and arrive at estimates comparable to those made with individual-level survey data for a non-English speaking context. With high-quality surveys continuing to be a highly popular method for estimating public opinion and the baseline for assessing the quality of (other types of) public opinion data (e.g., Daikeler et al., 2024; Sturgis & Luff, 2021), estimates based on LLMs should provide signals that are at least as good as surveys if they are to complement the latter.

Using the example of voting behavior, we provide a twofold methodological contribution to public opinion estimation using LLMs. We (1) show how a popular LLM performs in estimating voting behavior in a non-English context, Germany, compared to survey data, and (2) analyze which individual-level factors influence its predictions. Thus, we indirectly investigate how LLM performance varies across contexts that are less represented in the training data by evaluating its output for a less-represented context. Overall, in investigating the suitability of using LLMs for public opinion estimation in a new context, our study contributes to the growing body of research on the extent to which LLMs can be leveraged for research in the social sciences.

Background

Synthetic Samples

In survey research, synthesizing respondent samples has been argued to be one especially relevant application of LLMs. Such samples might allow for pre-testing survey questions on different

population segments faster and cheaper. They might also potentially supplement—by counterbalancing unit- or item-nonresponse (e.g., Kim & Lee, 2023)—or, as some hope, even replace survey-based data collection and public opinion estimation based on human samples, for example, in the context of political polls estimating voting behavior. The underlying idea researchers leverage is that LLMs are based on human-created data and might therefore potentially reflect humans' underlying attitudes and behaviors.

Trained on vast amounts of text data, LLMs generate a conditional probability distribution of how likely given tokens, that is, particles of words, are followed by specific other tokens. Presented with a string of words (*LLM input*), LLMs then draw on this probability distribution to predict words that are likely to follow (*LLM output*). For example, given the input "In the 2020 US presidential elections, I voted for," LLMs are more likely to complete the sentence with "the Democratic candidate" or "the Republican candidate" than with other terms unrelated to candidates or parties. The sentence is more or less likely to be completed with either vote choice depending on the training data, the configuration of the LLM algorithm, as well as any other information provided as input. LLMs are based on large, selected corpora of internet-sourced data, such as selected websites, book collections, and social media data, for example Reddit data from selected subreddits (see, e.g., Brown et al., 2020). As this training data potentially includes factual, attitudinal, and behavioral data about people, LLMs have been argued to provide a novel method for estimating public opinion in a population by creating synthetic samples: LLMs can be prompted repeatedly to answer survey questions, mimicking human respondents by providing individual-level characteristics as input. The distribution of responses provided in the output might serve as an estimate of the population. However, as of yet, widely-used LLMs do not learn from new data in real-time, but instead are trained on historical data up to a certain time point (see, e.g., OpenAI, n.d.). Therefore, these LLMs cannot take into account new information on current events that might influence public opinion.

Several recent studies have investigated the potential use of LLMs for replicating human samples in public opinion research, particularly in the area of political polling. For example, Argyle et al. (2023) prompted GPT-3 to respond to survey questions from the American National Election Study (ANES), reflecting different demographic subgroups of the population. The study found that the LLM-generated responses, on aggregate, closely matched the actual responses in the ANES data, and suggests that LLMs might even be able to estimate public opinion and voting behavior for time points exceeding their own training data. Similarly, Chu et al. (2023) showed that BERT, when trained on news media data, can emulate the attitudes of US subpopulations who consumed news media. Benchmarking the LLM responses against distributions from several surveys by Pew Research Center and the University of Michigan, their findings are robust to prompt wording and variation in media input. Other studies, however, have come to conflicting conclusions. For example, having GPT-3.5 impersonate ANES respondents and answer a set of survey items, the results by Bisbee et al. (2024) were mixed. While the average item scores produced by the LLM were similar to those obtained from the survey data, the LLM-based results had a smaller variance and resulted in different coefficients when regressing the prompt variables on the response. Furthermore, the responses were not robust to prompt wording and across time. Dominguez-Olmedo et al. (2023) had a large range of different language models respond to an entire questionnaire, benchmarking against the American Community Survey. In this study, however, even the aggregate estimates derived from the LLM responses did not match those of the human population. Finally, Santurkar et al. (2023), using the American Trends Panel survey, discovered substantial misalignments for specific subgroups. Testing several LLMs' "default" responses, not providing any further contextual information, as well as responses when prompting the LLMs to impersonate certain subgroups, the authors concluded that LLM-based samples cannot replicate human samples.

Challenges in Generalizability

A limitation of these existing studies is that they almost exclusively focus on the US population. To better understand if and under which conditions LLMs can be used for public opinion research, it is crucial to assess whether they can be applied for research in other national contexts. Several factors might limit the generalizability of previous findings beyond the United States.

Country-Level Factors. LLMs are likely better able to emulate public opinion for the United States than for other countries due to country-level factors associated with the training data. First, since LLMs are trained on text data from the internet, the amount of available native-language training data for developing LLMs is considerably smaller for any country with a native language other than English. For example, less than five percent of the content on the internet is estimated to be German, compared to English with over 50% (W3Techs, 2024). It is unclear how LLMs transfer their "knowledge" between training data in different languages and what "knowledge" is accessed when prompted in English about a non-English speaking population (see, e.g., Lai et al., 2023; Nie, Shao, et al., 2024, Nie, Yuan, et al., 2024). In either of these two processes, native, potentially more authentic, "knowledge" risks being underrepresented if LLMs are only accessing English-language training data.

Second, a country's societal and political structures may differentially affect the determinants of public opinion. These idiosyncratic relationships may not be sufficiently represented in LLM training data. For example, Argyle et al. (2023) showed that GPT-3 mirrored the relationships between subgroup characteristics and voting behavior in the US two-party system. It is unclear, however, whether these findings can be extended to multi-party systems, where the dynamics of voting behavior can follow fundamentally different patterns (Campbell et al., 1960; Lazarsfeld et al., 1944) due to (a) the number of parties, (b) issue alignment, and (c) strategic voting.

(a) Predicting voting behavior in multi-party parliamentary democracies is inherently more difficult than predictions for the two-party, first-past-the-post presidential democracy of the United States. Statistically, at a very basic level, the probability of making a correct prediction is inversely proportional to the number of parties competing.
(b) Moreover, the higher complexity of multi-party systems, also in terms of more potential combinations of issue positions, makes the voting decision more complex for voters. The clear binary alignment of certain issue positions is not obvious outside of the United States. Additionally, different social structures can lead to different policy-issue salience and conflicts. When using information about demographic or attitudinal subgroups to infer voting behavior without having been trained in these differences, LLMs are thus likely to wrongly project the more prominent political cleavages of the United States onto other contexts.
(c) Finally, in many multi-party systems, proportional representation and minimum thresholds create voters who vote strategically. These complex decision-making processes are often made spontaneously, in response to parties' popularities in current polls and the specific voting district, and therefore not explicitly discussed online. The concept of "swing voters" therefore is slightly different from that of the United States, as it is simply more common for voters to switch parties depending on the context (regarding policy issues and party popularity) in which the election takes place.

Not the least because it is usually the more politically interested and polarized who post on the internet (e.g., Kim et al., 2021; Muhlberger, 2003; Tucker et al., 2018), internet discussions, however, often tend to conflate political complexities to two camps (e.g., Yarchi et al., 2021). It is therefore likely that LLMs cannot mirror the more complex decision-making process in multi-party systems given the available training data.

The Digital Divide. Relatedly, the training data is likely affected by coverage bias. The difference between the general population and the population of internet users, the so-called "digital divide" (e.g., Lutz, 2019), may impact how representative the training data is of the population (see e.g., Clemmensen & Kjærsgaard, 2023). For example, the socio-demographic digital divide in Germany is slightly different from that in the United States (see Schumacher & Kent, 2020). As the composition of the online and offline populations differs between regions and countries (see also International Telecommunication, 2022), a country's societal structure may affect the bias in the LLM training data used to estimate public opinion.

In addition, there may be structural and attitudinal differences related to how people in a given society use the internet, that is, between those who actively produce or contribute to the text captured and more passive internet users in general, and between the authors of texts selected for training LLMs and other internet users specifically. For instance, the training data for GPT-3 is not a random sample of internet text, but heavily relies on very few sources, including Wikipedia, Reddit, and two collections of books (Brown et al., 2020)—sources that generally tend to be authored by rather homogenous communities: For example, Wikipedia reports that a plurality (20%) of its editors reside in the United States, edit the English Wikipedia (76%), and that, among editors of the English Wikipedia, 84% are male (Wikipedia, 2024; c.f. Hill & Shaw, 2013). Overall, the "knowledge sources" of LLMs are heavily concentrated on the English-speaking US context, which is then reflected in their outputs (Johnson et al., 2022).

These factors converge in what can be described as a "black box" of LLMs' internal workings. In this paper, we seek to empirically assess whether or not previous findings regarding public opinion estimation with LLMs can be generalized in the first place, not why they are (not) generalizable. Not only is empirically testing the latter contingent on the former, it would also require a broader scope and insights into the LLM "black box" that the research community does not currently have.

Comparative Research

Although there has been some cross-national and cross-lingual research on attitudinal biases of LLMs, these studies either did not explicitly estimate public opinion in general or did not do so for different population subgroups. For example, Motoki et al. (2023) and Hartmann et al. (2023) found that GPT's default political orientation was biased towards left or progressive ideologies in several two- and multi-party systems. Prompting ChatGPT with political questions that can be mapped onto ideological coordinates, Motoki et al. (2023) compared its responses given without any context to those it gave impersonating a partisan and found that the context-less default was more similar to the left partisan. However, the authors did not investigate the individual attitudes of the general public, but instead showcased what GPT "believes" a-priori partisans' political ideology to be (Motoki et al., 2023) or extrapolated from ChatGPT's responses to voting advice application questions to its likely vote choice (Hartmann et al., 2023). Durmus et al.'s (2023) cross-national study is closer to the synthetic-sample approach. The authors tested a custom LLM on entire questionnaires, both its default and when impersonating people from different countries. When comparing the LLM responses to several cross-national survey datasets (Pew Global Attitudes and the World Values Survey), they found that the LLM default responses tended to be more similar to the American and European benchmark data and reflected harmful country-level stereotypes for the other countries. Translations to a country's target language did not always improve the LLM responses' similarity to its speakers' attitudes. But while Durmus et al. (2023) compared English to Russian, Chinese, and Turkish prompting, the authors only used generic country personas ("How would someone from [country] answer this question?"), without considering specific subgroups, allowing only for aggregate cross-country comparisons. Bisbee et al. (2024) also conducted a cross-national test and found that ChatGPT's performance was similarly poor across countries for several survey items on public opinion, with a tendency to predict attitudes that are

more common in the benchmark survey data. Contrary to the authors' expectations, the accuracy for the United States was among the lowest. The authors only used prompts in English and did not test how the prompt language is related to performance. However, research suggests that input language impacts output quality (e.g., von der Heyde et al., 2025; Li et al., 2024), with prompting in non-English languages resulting in more US-centric output nevertheless (e.g., Durmus et al., 2023; Havaldar et al., 2023; Johnson et al., 2022; Wang et al., 2023). Thus, it remains unclear to what extent LLMs can be used for estimating individual-level public opinion outside the much-researched, two-party, English-dominated context of the United States, especially when using prompting languages other than English, and especially for smaller linguistic populations such as Germany.

The Case of Germany

In our study, we assess LLMs' suitability for estimating public opinion in Germany by focusing on voting behavior, which is a frequently studied outcome of interest in public opinion research. Germany serves as an example of a Western European democracy, with public opinion formed in the context of not two, but several political parties. Germany has a parliamentary electoral system with proportional representation and its multi-party system is currently characterized by six parties (Schmitt-Beck et al., 2022b): the center-right Christian conservatives (CDU/CSU), the center-left Social Democrats (SPD), the right-of-center, conservative-liberal Free Democrats (FDP), the left-of-center, environmentalist Green party (Greens), the Left party, and, more recently, the far-right "protest" party "Alternative for Germany" (AfD).[1] Moreover, it is an example of a country using a language not as dominant in online discourse as English but still relevant enough to allow for testing of our training data-related arguments, that is, differences in country-level factors and coverage biases affecting the training data. In what can be considered a "next-best" case scenario for LLM-based public opinion estimation, Germany presents a middle ground between the United States and other societies which are represented in the training data even less, which might pose a challenge for testing synthetic sampling. Findings in LLM-based public opinion estimation for Germany can be informative for countries with similar characteristics, and even those more underrepresented in the training data in terms of language and society: detecting limitations in LLMs' ability to estimate public opinion in this context would make it likely that this ability is even more limited in more structurally complex, under-researched, or underrepresented contexts.

The social structures dividing the German electorate differ substantially from those characterizing the United States (see, e.g., Lipset & Rokkan, 1967; Brooks et al., 2006; Ford & Jennings, 2020; Sass & Kuhnle, 2023). Moreover, the determinants of voting behavior on the micro-level play out in a different way than in the US context: Partisanship and traditional socio-economic and religious cleavages and their impact on voting behavior have declined (Berglund et al., 2005; Dalton, 2014; Franklin et al., 2004; Jansen et al., 2013; Elff & Rossteutscher, 2011; Schmitt-Beck et al., 2022a, 2022b). At the same time, the socio-cultural dimension (Inglehart, 1977; Schmitt-Beck et al., 2022) has become more important for voting behavior (Dalton, 2018). As a result of these developments, there are signs of situational issue-voting (Schoen et al., 2017) based on current salient and divisive topics, such as immigration (e.g., Kriesi et al., 2006).

Data and Methods

Data Collection

Benchmark Data and LLM Selection. In order to examine to what extent LLMs can estimate public opinion in Germany, we simulate a sample of eligible voters in Germany using GPT-3.5. We echo existing research designs in benchmarking the LLM's predicted vote choices against those reported by the survey respondents in the German Longitudinal Election Study (GLES, see Appendix I for details). While

surveys are not free from errors, they are currently the best available data source on public opinion on the individual level, allowing us to assess LLM performance for different subgroups of the population.

To ensure comparability with previous studies (Argyle et al., 2023; Bisbee et al., 2024; Dominguez-Olmedo et al., 2023; Hartmann et al., 2023; Motoki et al., 2023; Santurkar et al., 2023), we rely on GPT, which also has the advantage of being one of the largest language models available and being broadly accessible, making it a likely choice for potential applications in academia, industry, and by the public. We choose the 2017 German general election because it definitely occurred before the training data cutoff for our specific LLM in June 2021 (OpenAI, 2024b), with information about the election's context thereby likely included in the training data. If we find limitations in GPT's ability for estimating voting behavior for an election that occurred within the range of its training data, we cannot expect the LLM to perform well in predicting public opinion in contexts beyond its training data.

Prompt Creation. For the prompts provided to GPT-3.5, we create personas individually simulating each of the 1905 voting-eligible participants in the 2017 post-election cross-section of the GLES who reported their vote choice GLES (2019). The personas include individual-level information on 13 of the most common factors associated with voting behavior as identified in the literature about electoral behavior in Germany (c.f. Schmitt-Beck et al., 2022a, 2022b; Schoen et al., 2017; Klein, 2014). These variables comprise age, gender, educational attainment, income, employment status, residence in East/West Germany, religiosity, ideological left-right self-placement, (strength of) political partisanship, attitude towards immigration, and attitude towards income inequality.[2] Missing values on any of the variables are imputed for n = 377 respondents (20% of the sample) using multivariate imputation by chained equations (van Buuren & Groothuis-Oudshoorn, 2011). As a robustness check, we adjust the prompt using only the non-imputed variables for the respondents with missing values and compare the results (see Appendix X). We then feed these personas as prompts to GPT-3.5 in German, using the completions-API,[3] alongside the request to complete the last sentence with the respective person's vote choice in the 2017 German parliamentary elections. An example prompt is shown in Figure 1, translated to English for illustrative purposes (see Appendix II for the German original).

We choose to prompt the LLM in German because the aim of our study is to examine the usability of LLM-generated synthetic samples for public opinion estimation in a non-US- and/or -English context, in order to inform potential applications outside of the US. Not all local public opinion items are available in English with a faithful translation and testing of concepts. From a normative point of view, requiring an instrument to be translated to English for LLMs to be usable is questionable, as it risks further marginalizing other languages—also when considering LLMs learn from their interactions with human input. Indeed, it is unclear whether English-language prompting would yield better results due to the larger amount of training data. As we have argued, one could conversely expect an LLM to more closely approximate attitudes in the target population when prompted to access those probabilities it has learned from native-language training data, as these may be more likely to represent "authentic" attitudes. However, as native-language training data is unequally distributed across target populations, we expect these approximations to be comparatively worse than for a target population whose native language is English (see also Durmus et al., 2023). We leave a comparison of results when using English versus native-language prompting to future research, as it would be out of the scope of the current paper.

LLM Configuration. Based on the outputs of a pilot test (see Appendix III for details), we calibrate GPT-3.5's text-davinci-003 to a temperature of 0.9 and a response length of maximum 30 tokens.[4] We choose a high temperature to be in line with similar studies (e.g., Argyle et al., 2023; Bisbee et al., 2024) and to simulate the non-determinism in human responses to survey

*I am **28** years old and **female**. I have **a college** degree, a **medium** monthly net household income, and am **working**. I am **not religious**. Ideologically, I am **leaning center-left**. I **rather weakly** identify with **the Green party**. I live in **West** Germany. I think the government should **facilitate** immigration and **take measures to reduce** income disparities.*
Did I vote in the 2017 German parliamentary elections and if so, which party did I vote for? I [INSERT]

Figure 1. Example prompt (translated from German, variables bolded for emphasis).

questions (e.g., Zaller, 1992): Since (reported) human voting behavior is not deterministic, forcing the LLM to always pick the same or most probable option by setting the temperature to zero would not be representative of human behavior.[5] We collect our data in July 2023 (main sample) and November 2023 (robustness checks). Since the release of GPT-3.5 and its API, OpenAI has performed several changes to both the language model and its data accessibility, including deprecating the possibility of storing token probabilities via the API, that is, the probability with which a sentence is completed with the selected completion token. However, research suggests that first-token probabilities do not always match completions when prompting an LLM with survey questions, especially for sensitive topics that are more likely to induce a refusal from the LLM (Wang et al., 2024). First-tokens also are more sensitive to the prompt format than text output. These limitations make first-token probabilities an infeasible evaluation metric. To nevertheless account for the probabilistic nature of GPT's responses beyond a single text completion, we adopt procedures established in multiple imputation (Rubin, 1987). Specifically, we sample five completions per persona and estimate the variance between these samples.[6] By using multiple completions, we can investigate the range and variability of GPT's outputs. This variance analysis helps us grasp the model's behavior and the reliability of its responses, providing insights into the consistency and robustness of the model's text generation. This way, we account for both human (temperature) and LLM (number of samples) randomness in our estimates. Our data thus includes 9525 LLM-generated completions.

Vote Choice Extraction. We then extract the party names from the LLM completions as defined by a set of accepted keywords per party (see Appendix IV), also considering non-voters and invalid votes. 1427 completions initially did not contain a vote choice. For these, we re-prompt the LLM up to two times, replacing the respective initial completion, resulting in 87 or 0.9% of final completions not containing a vote choice (see Appendix V for details and Appendix X for an investigation of systematic patterns in these personas/completions).

Analysis

We compare the survey-reported and LLM-generated vote choices to investigate the extent to which the responses differ in terms of vote choice as well as how the two data sources weigh the prompt variables in estimating vote choice. This approach allows us to not only assess whether GPT-3.5 is able to estimate the voting behavior of the German general population on aggregate, but also whether it can make equally accurate estimates for different population subgroups.

To tackle our first research question, we compare the aggregate distribution of vote shares across parties according to GPT-3.5 to that based on GLES data. We also estimate multinomial

regression models of voting behavior as reported in GLES and predicted by GPT-3.5, respectively, on the prompting variables. These models serve two purposes: Relating to our first research question, we evaluate GPT-3.5's predictive performance by comparing its predictions to the predicted values of the GLES-based regression model. We do this by calculating precision, recall, and macro F1 scores[7] overall and per party, for both the LLM-based predictions and the GLES model predictions. Of course, perfect predictions of (reported) individual voting behavior are unlikely, due to the limited predictability of any election. Therefore, we also test whether the LLM at least mirrors the survey data's correlates of voting behavior, by comparing the regression models in terms of effects of specific individual characteristics. This addresses our second research question.

For estimating the regression models, we fit maximum conditional likelihood models based on a neural network with a single hidden layer (Venables & Ripley, 2002). For all regression models, we exclude 78 respondents for whom at least one of the five GPT samples did not contain an explicit vote choice to ensure comparability across samples,[8] and treat ordinal independent variables with at least five categories as numeric. In order to obtain just one estimate from the five GPT samples, we employ variance estimation as established in multiple imputation research (Rubin, 1987). For each analytical method, we calculate each estimate separately for each sample, and then aggregate across the five samples to obtain the average estimate and total standard error. For example, for our regression models, we run five separate regressions, one per sample, and compute the average coefficient and standard error (as $\sqrt{\mu(SE_{\beta}^2) + 1.2\sigma_{\beta}^2}$) (Rubin, 1987) to construct confidence intervals.

All analyses are conducted using the software R (R Core Team, 2024), version 4.3.0, especially the packages *tidyverse* (Wickham et al., 2019), *mice* (van Buuren & Groothuis-Oudshoorn, 2011), *rgpt3* (Kleinberg, 2023), *nnet* (Venables & Ripley, 2002), and *marginaleffects* (Arel-Bundock, 2021).

Results

RQ1: Do LLM-Based Samples Provide Similar Estimates of Voting Behavior as National Election Studies?

Distribution of Vote Choice Across Parties. On aggregate, the GPT-based distribution of vote shares across parties differs markedly from that of the national election poll. Compared to the GLES sample, GPT-3.5 overestimates the share of Green, Left, and non-voters, while underestimating the share of FDP and AfD voters as well as voters of small parties (see Figure 2). For the two major parties, CDU/CSU and SPD, there are no significant differences.

Predictive Performance. Across the five samples drawn from GPT, there is very little variance in terms of whether the GPT prediction matches the vote choice individual respondents reported to GLES. On average, only 39% of GPT-3.5's predictions match the survey data. The F1 scores indicating the LLM's predictive accuracy are best for the CDU/CSU (0.6), followed by SPD and Greens, and much worse for FDP and AfD (around 0.3; see Appendix VII). Comparing these scores to those of predictions based on a multinomial model fitting *GLES-reported* vote choice on the prompt variables, the GLES model creates better predictions than GPT. Given the same demographic and attitudinal information, the GLES model consistently performs better, both overall (macro F1 0.39 vs. 0.52) and with regard to specific parties, and most notably for the AfD and FDP (both above 0.5; see Appendix VII). These differences are informative for both the aggregate and subgroup analyses. For the aggregate, they confirm what the overall distribution of

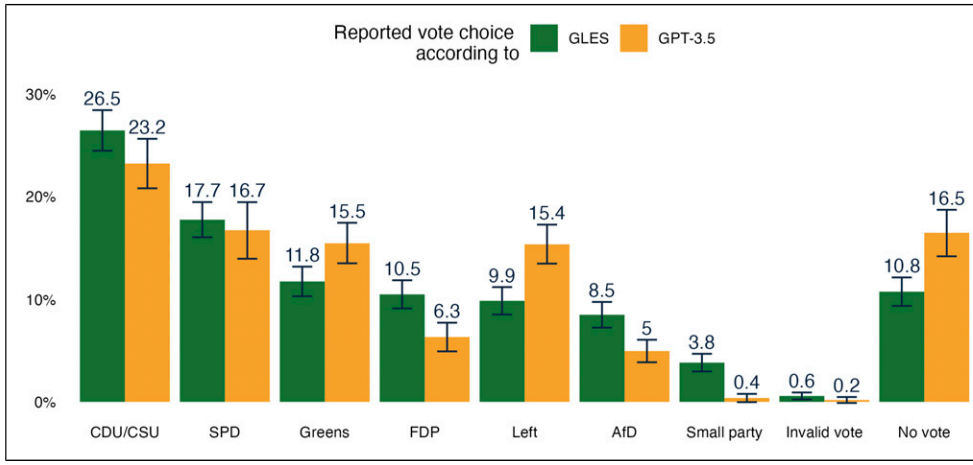

Figure 2. Distribution of vote shares as estimated by GLES and GPT.

vote shares indicated: The LLM-based estimates of voting behavior are different from survey-based ones, as it is easier for GPT-3.5 to predict voters of Germany's center- and left-leaning parties than right-leaning ones. For subgroups, the higher predictive accuracy of the GLES model justifies benchmarking the predictors of voting behavior according to GPT-3.5 against those in the GLES model.

RQ2: How do LLMs' Estimates of Voting Behavior Deviate from National Election Studies for Different Subgroups of the Population?

Comparing the impact the prompting variables have on actual GPT- versus GLES-reported vote choice in multinomial regressions (see Figure 3 for a visualization of the differences in effect size accounting for effect direction, and Appendix VIII for the full tables), the models show that GPT-3.5's predictions of vote choice are reliant on certain cues in the prompts, which often do not match the effects the survey data indicates.[9] The model indicates that GPT-3.5 appears to be taking partisanship into particular account when asked to predict people's vote choice. For example, as shown in Figure 4, GPT-3.5 exhibits similar positive effects as the GLES model for SPD, Green, FDP, Left, and AfD partisans on the probability of voting for the respective party. Likewise, it picks up the signal of left-right ideology for the extremes of the party spectrum: the Left and AfD. However, apart from the far-right and -left, GPT-3.5 does not mirror GLES when it comes to the importance of ideology, for example, on voting for the CDU/CSU or the Greens. Moreover, when it comes to partisanship, it does not account for *negative* partisanship: For example, the GLES data suggests a systematic underlying pattern between Green and AfD voters—Green partisans are significantly *less* likely to vote for the AfD, and vice versa. In sum, while partisanship and ideology are important factors influencing voting behavior, GPT-3.5 only picks up on broad trends, without regards for more complex dynamics.

Its dominant reliance on party identification as a predictor of vote choice can help explain why GPT-3.5 underestimates the vote shares for FDP and AfD, as observed in the previous subsection: Although most partisans indeed vote for the party they identify with (see Table A8 in Appendix IX), only half of the voters of the FDP, AfD, and small parties also identify with their chosen party. Thus, in presence of a partisan cue, GPT-3.5 predicts partisans to vote in line with their party identification. For voters without this cue, its predictions falter.

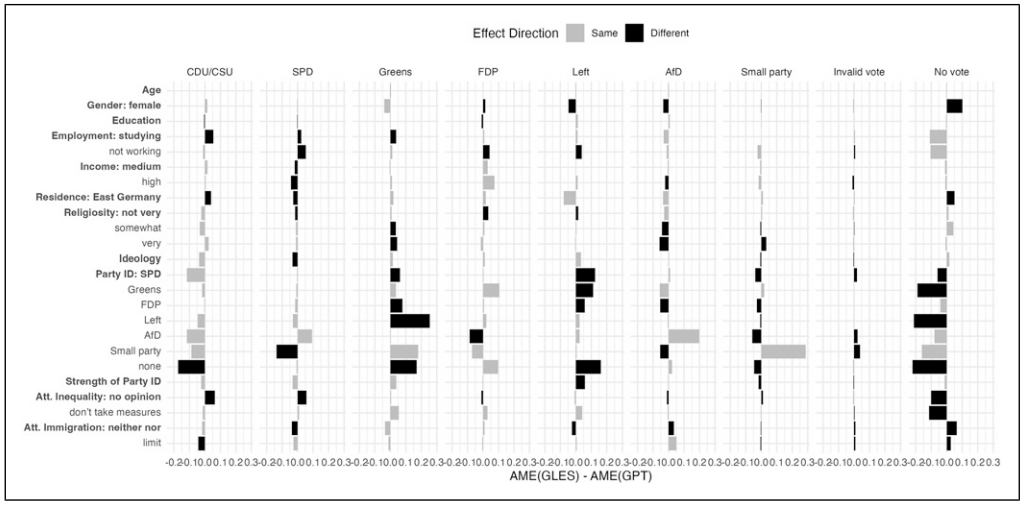

Figure 3. Difference in average marginal effects of prompt variables on vote choice, as estimated by GLES and GPT. *Note*. Average marginal effects describe the average of the fitted results of the model after first making individual predictions for each row in the original dataset, mirroring the real data (c.f. Heiss, 2022). Difference denotes subtraction of GPT-based AME from GLES-based AME. Effect direction refers to positive or negative effects for GLES and GPT. Example: Black negative bars denote cases where the effect estimated by GPT is positive, while that estimated by GLES is negative. Grey negative bars denote cases where both effects are negative, but that estimated by GPT is (closer to) zero. Vice versa for positive bars. For reference categories, see Table A7.

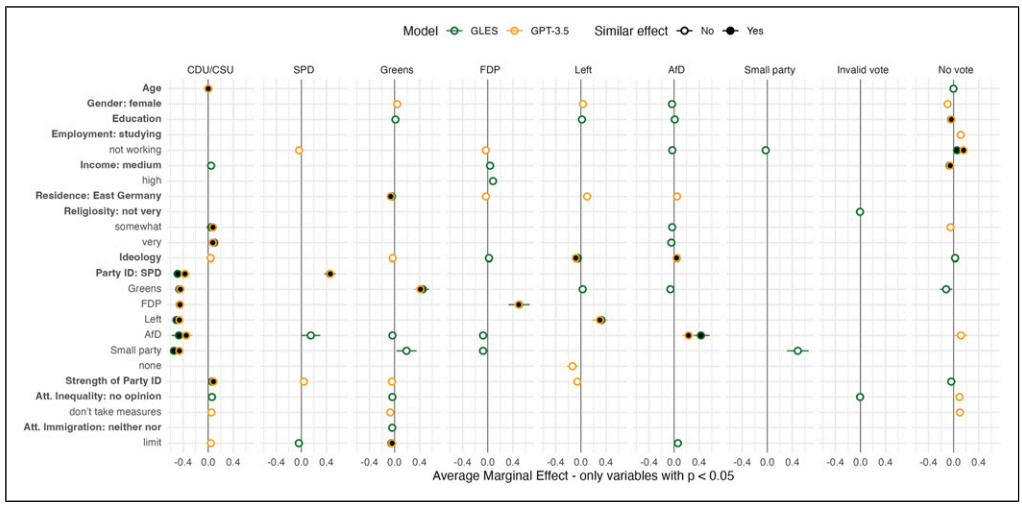

Figure 4. Average marginal effects of prompt variables on vote choice, as estimated by GLES and GPT. Note: Average marginal effects describe the average of the fitted results of the model after first making individual predictions for each row in the original dataset, mirroring the real data (c.f. Heiss, 2022). "Similar effect" denotes effects that are significant for both the GLES and GPT model and point in the same direction (positive or negative) regardless of magnitude. For reference categories, see Table A7.

However, overall, there are more differences than similarities in predictors of vote choice between the LLM-generated and survey data. For the remaining attitudinal variables as well as most demographic indicators, GPT-3.5's predictions assume different mechanisms than what the GLES data suggests, following general patterns identified by previous research on German voting behavior, but not considering the nuances of more complex subgroups. For example, the GPT model, but not the GLES model, suggests residents of East Germany are more likely to vote for the Left or AfD, females are more likely to vote for the Greens or Left, and non-workers less likely to vote for the SPD or FDP (which traditionally have catered to different segments of the working population). Contrary to the GLES model, it does not consider education and income as important factors for distinguishing the likelihood of voting for the Left, Greens, or FDP versus the AfD, nor the importance of religiosity for distinguishing CDU/CSU from AfD voters.

Similar to what can be observed for (negative) partisanship, GPT-3.5 does not capture the complex effects of attitudes towards inequality and immigration. For example, while the GPT data matches the GLES data in indicating that wanting to limit immigration *decreases* the likelihood of voting for the Greens, the GLES data also indicates that such an attitude *increases* the likelihood of voting for the AfD.

All in all, when considering survey data as ground truth, voting behavior in Germany depends on a different number and kind of factors than GPT-3.5's predictions would suggest. GPT-3.5 bases its predictions on partisanship as well as indicators for common subgroups of voters for a specific party. This finding suggests that GPT-3.5 relies on rather simplified signals in making its predictions, without necessarily considering other, more complex mechanisms in the individual voting decision-making process.

Discussion

Our study assessed the capabilities of a popular large language model (GPT-3.5 text-davinci-003) in estimating voting behavior for the 2017 German federal elections, using the reported vote choices from the respondents of the German Longitudinal Election Study (GLES) data as a benchmark. We created personas simulating every individual respondent in the GLES study. Prompts generated from these personas were then fed to GPT-3.5 via the OpenAI API with a request to complete the personas' vote choice. We compared GPT-3.5's predicted vote choices to respondents' actual vote choices for multiple political parties. Moreover, we conducted a focused subgroup analysis and compared the determinants of voting behavior for the GLES responses with those of the GPT predictions.

Using Germany as an example, we have shown that using LLMs for estimating public opinion in a similar way to surveys cannot simply be generalized beyond the initial applications in the English-speaking context of the United States. In our findings, GPT-3.5 overestimated the survey-reported vote shares for the Greens, the Left, and non-voters by a significant margin, while it underestimated the vote shares for FDP and AfD when compared to GLES. The LLM's overall predictive accuracy was modest, with a macro F1 score of 0.39. It was notably more accurate for voters of the Greens, CDU/CSU, and the Left, but displayed poor predictive power for FDP and AfD voters.

In terms of determining factors that influence voting behavior, GPT-3.5's predictions largely hinged on straightforward indicators, such as strong party identification or ideology. However, when compared to the GLES data, it became evident that GPT-3.5 deviated substantially on more complex variables, like attitudes towards immigration or economic policy, socio-demographic variables, or the particular dynamics of partisanship. This discrepancy in correlates suggests that while GPT-3.5 might capture broad-brush trends tied to partisanship, it tends to miss out on the nuanced, multifaceted factors that sway individual voter choices, thereby limiting its predictive

accuracy. As a consequence, relying on LLM-based estimates does not help researchers when predicting voting behavior: Partisans are typically easy to predict as long as they vote in line with their party identification. However, if information on partisanship is necessary for GPT-3.5 to make a prediction, and it cannot evaluate other, more complex relationships in absence of this information, then not only are LLM-based samples not helpful in predicting how non-partisans, weak partisans, or "inconsistent" partisans—all groups who likely swing an election—will vote. They also risk underestimating vote shares for parties with fewer (reported) partisans. Moreover, the absence of mirroring *negative* relationships between factors such as partisanship and immigration attitudes could lead to an underestimation of the popularity of certain parties when applying LLM-based sampling to estimate public opinion. In our case, GPT-3.5 modeled decreasing likelihoods of certain individuals voting for the Greens, without a correct indication of who these individuals would be more likely to vote for (in this case, the AfD), which, consequently, got underestimated.

Naturally, predicting voting behavior in a multi-party system is inherently more difficult than in a two-party system. This challenge remains when transferring the task to LLMs, and therefore is likely one of the reasons why we cannot expect LLMs to work similarly well in all contexts. Moreover, public opinion in general, and voting behavior in particular, is situated in the positionalities and temporalities (including the susceptibility to shock events) of the individual. Therefore, neither survey nor LLM data will produce 100% accuracy. While researchers hope that LLMs can help them uncover patterns and make predictions where traditional methods struggle, our study underscores the limited applicability of LLM-based synthetic samples. This difficulty is compounded by differences in social structures leading to differential issue conflicts, and by limited nuanced, native language, and target-population-representing internet text from which LLMs could learn about these complexities. As McCoy et al. (2023) demonstrate, LLMs may fail at one task while being successful at another task of comparable complexity, if the former is less commonly represented in the training data.

Thus, considering the types of text data that were used to train GPT may shed light on its predictive limitations. GPT is trained on a large, but mainstream and not necessarily diverse corpus of text data that includes a selection of websites, books, and other publicly available texts (Brown et al., 2020). As a result, the LLM may be predisposed to make predictions based on generalized or commonly represented political beliefs and more typical, well-researched voter groups, hence struggling with accurately predicting the behavior of voters for the AfD and other non-conforming groups.

This finding underscores the limitation of applying GPT to electoral predictions without accounting for the biases and limitations inherent in its training data. It reaffirms that while GPT can provide certain broad insights, it may not be reliable for nuanced, subgroup-specific political predictions. Because such nuanced relationships are not mirrored, we prefer following a conservative interpretation of our results, although LLMs can cover general trends. This interpretation is in line with previous issues identified with generative artificial intelligence, which in the context of image generation have been found to reproduce and amplify oftentimes harmful stereotypes and biases (Bianchi et al., 2023; Nicoletti & Bass, 2023; Turk, 2023). Ultimately, using LLMs for estimating public opinion risks reinforcing existing biases. Placing our results in the discourse of existing findings, it remains questionable whether LLM-synthetic samples may be useful for public opinion research—both inside and outside the US. Indeed, even studies considering the US come to diverging conclusions (c.f. Argyle et al., 2023; Chu et al., 2023 vs. Bisbee et al., 2024; Dominguez-Olmedo et al., 2023; Santurkar et al., 2023). It thus appears that, similar to surveys conducted with non-probability samples, LLM-based synthetic samples *can* get it right sometimes, but not reliably so. Considering that LLM responses do not represent latent attitudes of an existing target population, but a probability distribution of most-likely next words, even the

validity of such measurements may be questioned. This notion is supported by evidence that public opinion data based on LLM-synthetic samples exhibits less variance than human data (e.g., Bisbee et al., 2024; Dominguez-Olmedo et al., 2023), and that this variance depends on probability sampling procedure, the prompt variables, and the LLM used (Roth, 2024). It may be argued that LLMs can still be considered useful despite these shortcomings as long as they provide insights which surveys cannot do. However, as our as well as other studies have shown, they do not meet this expectation so far.

Limitations

Our study encountered several limitations that should be considered when interpreting the results and offer avenues for improvement in future research. First, our study did not experiment with prompt design. We did not test the effect of variable ordering in the prompt on the predictions, which was beyond the scope of our study but could have potentially affected the results (Bisbee et al., 2024). Future research could engage with work on prompt engineering to optimize GPT's predictions. For example, future studies should specify the reference year for time-sensitive variables if it differs from the prompting year (such as age when applied to voting behavior), as Bisbee et al. (2024) suggest. Additionally, while our selection of prompt variables was rooted in existing research on voting behavior in Germany, we acknowledge that other factors might contribute to the voting decision-making process and thus could enhance the predictive power of the LLM. Furthermore, it is unclear how GPT draws on training data in another language than the prompt and completion language.

Second, recording token probabilities for the vote choice estimation through the OpenAI API (Argyle et al., 2023) was no longer possible at the time of data collection. This constraint highlights the dependency on the functionalities that API providers offer.

Third, the study used text-davinci-003 for its analyses, which may be less efficient and precise than newer language models. However, this choice was made at the time of writing to ensure comparability with existing studies and due to the API availability. The constant "under the hood" changes to and rapid advancement of these language models and their APIs, with the text-davinci-003 model used in this research deprecated in January 2024 (OpenAI, 2024a), raises concerns about the replicability of research such as ours (e.g., Spirling, 2023) and challenges social scientists to continuously re-evaluate previous findings. Our study thus can serve as a reference point for understanding the evolution of LLMs in the realm of cross-cultural public opinion estimation. Newer GPT versions with potentially better performance on this task have since emerged and should be tested, as should open-source LLMs. However, so far, the conclusions remain pessimistic when relying on off-the-shelf LLMs as opposed to fine-tuned ones (compare von der Heyde et al., 2025 to Ahnert et al., 2024; Holtdirk et al., 2024). Beyond the choice of LLM, results may vary depending on the benchmark survey or specific outcome measured. Investigating the influence of these factors falls outside the scope of this study and is recommended for future research.

Fourth, we recognize that our findings for Germany can at best be generalized to Western European socio-political contexts. We argued that our selected case presents a reasonable middle ground for assessing the suitability of LLMs for public opinion research, as it is distinguishable from the United States on the factors we identified as potentially limiting this suitability. While the limitations in LLM public opinion estimation we have found in the German context can be considered unpromising for more structurally complex, under-researched, or disadvantaged societies, such research should be explicitly conducted. However, benchmarking the LLM's responses against reliable individual-level public opinion survey data implies that studies such as ours can only soundly be conducted in countries that already have a good survey infrastructure. On

the other hand, while we treated survey data as ground truth, surveys themselves are not free of errors, but can suffer from errors related to sampling, coverage, measurement, and nonresponse. However, we reiterate that in this paper, we were not primarily interested in whether survey- or LLM-generated data is better at accurately predicting actual election outcomes. Both data sources have idiosyncratic error sources leading to differences between their estimates and the actual election results. Comparing errors across data sources would have presented an additional research question that would have been beyond the scope of this paper, but provides an opportunity for future work.

Outlook

This paper contributes to the rapidly growing field of computational social science using LLMs. Many other aspects and conditions under which LLMs might be used for public opinion research are yet to be explored. For example, researchers have suggested that LLMs might be helpful for estimating specific minoritized subgroups' attitudes, but this remains to be tested, for example, by benchmarking against special population surveys. Moreover, most existing studies have tested whether LLMs are able to "predict the past," that is, benchmarking against survey data from a time included in LLMs' training data. Future research should tackle the question of whether LLMs can predict future voting behavior based on past training data, for example by using pre-election panel survey data ahead of an upcoming election for the LLM input and comparing the LLM output to the post-election survey data after the election took place.

Extending the scope to other linguistic, socio-structural, and political contexts, comparative studies could employ cross-national individual-level benchmark datasets. Beyond further examining the contexts in which LLMs can(not) be used for public opinion estimation, such studies should systematically uncover which country-level factors drive this feasibility through multilevel or meta-analyses.

Finally, researchers could explore designing an LLM that is optimized for the purpose of survey research, drawing on comparative evaluations of existing LLMs' performance and the unique requirements of survey research. Fine-tuning LLMs on pertinent public opinion data (e.g., Ahnert et al., 2024; Holtdirk et al., 2024; Kim & Lee, 2023) is a first step in this direction.

Conclusion

We have shown that GPT-3.5 is not suitable for estimating voting behavior overall and across (sub) populations, as it exhibits algorithmic bias on two levels. From a *cross-sectional* perspective, although the LLM-generated data carried some signal that was able to account for the "big picture" of voting trends, it was unable to pick up on nuances of voter groups, thereby being biased against population subgroups not conforming to the mainstream. From a *cross-national* perspective, GPT-3.5's performance in estimating voting behavior was not as good for Germany as some comparable studies found for the United States. Even considering our interpretation of the results is rather conservative, predictive performance is likely to be even worse for countries, contexts, and populations who are reflected in the LLM training and alignment process even less. The application of large language models to public opinion estimation thus is limited to (sub)populations to which their training data is biased—whether this is due to contextual complexity or a lack of linguistic or digital representativity of other populations. More research is necessary to understand what exactly this bias in public opinion estimation depends on and how its sources interact. In sum, GPT-3.5 is better at estimating groups that dominate research and internet data—groups that researchers already know more about, only making LLM-based synthetic samples useful in very limited settings. Researchers need to be aware of these limitations when trying to apply large

language models in their work and take care not to reinforce existing biases. Only if large language models are equitable, just, and reflect the population's diversity in an unbiased manner may we be able to leverage them for estimating public opinion.

Acknowledgments

The authors would like to thank Max M. Lang, BSc, for his assistance in the automated processing of the LLM-generated data, and the FK^{2}RG research group for helpful feedback on earlier drafts of this paper.

Author Contributions

Leah von der Heyde conceptualized the research design, compiled and summarized the literature, collected, processed, and analyzed the data, and wrote large parts of the manuscript. Anna-Carolina Haensch conceptualized the research design, prepared the automation of data collection, wrote subsections of the manuscript, and provided comments and edits to the manuscript. Alexander Wenz refined the research design, prepared the data needed for data collection, wrote subsections of the manuscript, and provided comments and edits to the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethical Statement

Ethical Considerations

We use secondary data originally collected by GESIS – Leibniz Institute for the Social Sciences. For more information on informed consent as part of this data collection, see https://doi.org/10.4232/1.13235. No ethical approval was requested for the collection of data generated by large language models.

ORCID iDs

Leah von der Heyde https://orcid.org/0009-0002-8992-0251
Anna-Carolina Haensch https://orcid.org/0000-0001-6772-0393
Alexander Wenz https://orcid.org/0000-0002-4621-2418

Data Availability Statement

Data and code for the analyses conducted can be downloaded from https://github.com/leahvdh/Vox-Populi-Vox-AI

Supplemental Material

Supplemental material for this article is available online

Notes

1. At the time of writing, it is unclear whether the newly founded "BSW," which splintered from the Left party, will establish itself in the party system in the long term.
2. For details on the variables in the prompt, see Appendices I and II. For summary statistics, see Appendix VI.
3. For details, see Appendix III.
4. See Appendix III for detailed explanations.
5. Voting behavior is inherently situational. Therefore, perfectly capturing it with either survey or LLM data is unlikely. We discuss this in the Discussion section.
6. The primary purpose here is to explore the variance in GPT's responses, not to derive formal standard errors, as the assumptions for Rubin's rules, for example, are not fulfilled.
7. Scores range from 0 to 1, with higher values indicating better predictive performance. For a detailed explanation, see Appendix VII.
8. In Appendix X.3, we show that omitting 73 respondents who GPT classified as non-voters due to a (wrongful) ineligibility from the analysis does not change the results.
9. It is noticeable that aggregate estimates and individual-level predictive accuracy are better and more similar between GLES and GPT for parties where models of the two data sources share predictors, such as for the CDU/CSU.

References

Aaru. (n.d.). *Aaru: Rethinking the science of prediction*. https://aaruaaru.com

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2024). *Extracting affect aggregates from longitudinal social media data with temporal adapters for large language models* (No. arXiv:2409.17990). arXiv. https://arxiv.org/abs/2409.17990

Arel-Bundock, V. (2021). *Marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests* (p. 0.21.0) [Dataset]. https://doi.org/10.32614/CRAN.package.marginaleffects

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Supplementary material: Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 1–15. https://doi.org/10.1017/pan.2023.2

Ball, S., Kreuter, F., & Panickssery, N. (2024). *Understanding jailbreak success: A study of latent space dynamics in large language models* (arXiv: 2406.09289). arXiv. https://arxiv.org/abs/2406.09289

Berglund, F., Holmberg, S., Schmitt, H., & Thomassen, J. (2005). 5 party identification and party choice. In J. Thomassen (Ed.), *The European voter* (1st ed., pp. 106–124). Oxford University PressOxford. https://doi.org/10.1093/0199273219.003.0005

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In 2023 ACM conference on fairness, accountability, and transparency, Chicago, IL, USA, 12–15 June, 2023, pp. 1493–1504. https://doi.org/10.1145/3593013.3594095

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, *32*(4), 1–16. https://doi.org/10.1017/pan.2024.5

Brooks, C., Nieuwbeerta, P., & Manza, J. (2006). Cleavage-based voting behavior in cross-national perspective: Evidence from six postwar democracies. *Social Science Research*, *35*(1), 88–128. https://doi.org/10.1016/j.ssresearch.2004.06.005

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners* (arXiv: 2005.14165). arXiv. https://arxiv.org/abs/2005.14165

Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The American voter* (University of Michigan, Ed.; Unabridged ed). University of Chicago Press.

Chopra, F., & Haaland, I. (2023). Conducting qualitative interviews with AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4572954

Chu, E., Andreas, J., Ansolabehere, S., & Roy, D. (2023). *Language models trained on media diets can predict public opinion* (arXiv:2303.16779). arXiv. https://arxiv.org/abs/2303.16779

Chua, G. (2024, November 5). *An AI polling startup makes its predictions for the 2024 US election*. Semafor. https://www.semafor.com/article/11/04/2024/an-ai-polling-startup-polls-bots-predicts-harris-will-win

Clemmensen, L. H., & Kjærsgaard, R. D. (2023). *Data representativity for machine learning and AI systems* (arXiv:2203.04706). arXiv. https://arxiv.org/abs/2203.04706

Daikeler, J., Fröhling, L., Sen, I., Birkenmaier, L., Gummer, T., Schwalbach, J., Silber, H., Weiß, B., Weller, K., & Lechner, C. (2024). Assessing data quality in the age of digital social research: A systematic review. *Social Science Computer Review*, 1–37. https://doi.org/10.1177/08944393241245395

Dalton, R. J. (2014). Partisan dealignment and voting choice. In S. Padgett, W. E. Paterson, & R. Zohlnhöfer (Eds.), *Developments in German politics 4* (pp. 57–77). Macmillan Education UK. https://doi.org/10.1007/978-1-137-30164-2_4

Dalton, R. J. (2018). *Political realignment: Economics, culture, and electoral change* (Vol. *1*). Oxford University Press. https://doi.org/10.1093/oso/9780198830986.001.0001

Delve AI (nd). AI persona generator. https://www.delve.ai/ai-persona-generator

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). *Questioning the survey responses of large language models* (arXiv:2306.07951). arXiv. https://arxiv.org/abs/2306.07951

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., & Ganguli, D. (2023). *Towards measuring the representation of subjective global opinions in language models* (arXiv:2306.16388). arXiv. https://arxiv.org/abs/2306.16388

Elff, M., & Rossteutscher, S. (2011). Stability or decline? Class, religion and the vote in Germany. *German Politics*, *20*(1), 107–127. https://doi.org/10.1080/09644008.2011.554109

Ford, R., & Jennings, W. (2020). The changing cleavage politics of Western Europe. *Annual Review of Political Science*, *23*(1), 295–314. https://doi.org/10.1146/annurev-polisci-052217-104957

Franklin, M. N., Eijk, C. V. D., Evans, D., Fotos, M., Hirczy De Mino, W., Marsh, M., & Wessels, B. (2004). *Voter turnout and the dynamics of electoral competition in established democracies since 1945* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511616884

GLES. (2019). *Post-election cross section (GLES 2017) Nachwahl-Querschnitt (GLES 2017)* (4.0.1) [Dataset]. GESIS Data Archive. https://doi.org/10.4232/1.13235

Götz, F. M., Maertens, R., Loomba, S., & Van Der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, *29*(3), 494–518. https://doi.org/10.1037/met0000540

Hartmann, J., Schwenzow, J., & Witte, M. (2023). *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation* (arXiv:2301.01768). arXiv. https://arxiv.org/abs/2301.01768

Havaldar, S., Rai, S., Singhal, B., Liu, L., Guntuku, S. C., & Ungar, L. (2023). *Multilingual Language models are not multicultural: A case study in emotion* (No. arXiv:2307.01370). arXiv. https://arxiv.org/abs/2307.01370

Heiss, A. (2022, May 22). *Marginalia: A guide to figuring out what the heck marginal effects, marginal slopes, average marginal effects, marginal effects at the mean, and all these other marginal things are*. Andrew Heiss. https://www.andrewheiss.com/blog/2022/05/20/marginalia/#what-are-marginal-effects

Hernandez, I., & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, *76*(4), 1011–1035. https://doi.org/10.1111/peps.12543

Hill, B. M., & Shaw, A. (2013). The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS One*, *8*(6), Article e65782. https://doi.org/10.1371/journal.pone.0065782

Holtdirk, T., Assenmacher, D., Bleier, A., & Wagner, C. (2024). *Fine-tuning large language models to simulate German voting behaviour* (Working Paper). https://doi.org/10.31219/osf.io/udz28

Inglehart, R. (1977). *The silent revolution: Changing values and political styles among Western publics*. Princeton University Press.

International Telecommunication, Union. (2022, July 2). *Measuring digital development: Facts and figures 2022*. https://www.itu.int/hub/publication/d-ind-ict_mdd-2022/.

Jaimovitch-López, G., Ferri, C., Hernández-Orallo, J., Martínez-Plumed, F., & Ramírez-Quintana, M. J. (2023). Can language models automate data wrangling? *Machine Learning*, *112*(6), 2053–2082. https://doi.org/10.1007/s10994-022-06259-9

Jansen, G., Evans, G., & Graaf, N. D. D. (2013). Class voting and left–right party positions: A comparative study of 15 Western democracies, 1960–2005. *Social Science Research*, *42*(2), 376–400. https://doi.org/10.1016/j.ssresearch.2012.09.007

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). *The Ghost in the Machine has an American accent: Value conflict in GPT-3* (arXiv: 2203.07785). arXiv. https://arxiv.org/abs/2203.07785

Kim, J., & Lee, B. (2023). *AI-augmented surveys: Leveraging large language models and surveys for opinion prediction* (arXiv:2305.09620). arXiv. https://arxiv.org/abs/2305.09620

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

Klein, M. (2014). Gesellschaftliche Wertorientierungen, Wertewandel und Wählerverhalten. In J. W. Falter & H. Schoen (Eds.), *Handbuch Wahlforschung* (pp. 563–590). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-05164-8_13

Kleinberg, B. (2023). *rgpt3: Making requests from R to the GPT-3 API and ChatGPT*. R package version 0.4. https://github.com/ben-aaron188/rgpt3

Konstantis, K., Georgas, A., Faras, A., Georgas, K., & Tympas, A. (2023). Ethical considerations in working with ChatGPT on a questionnaire about the future of work with ChatGPT. *AI and Ethics*, *4*(4), 1335–1344. https://doi.org/10.1007/s43681-023-00312-6

Kriesi, H., Grande, E., Lachat, R., Dolezal, M., Bornschier, S., & Frey, T. (2006). Globalization and the transformation of the national political space: Six European countries compared. *European Journal of Political Research*, *45*(6), 921–956. https://doi.org/10.1111/j.1475-6765.2006.00644.x

Lai, V., Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. (2023). ChatGPT beyond English: Towards a Comprehensive evaluation of large language models in multilingual learning. Findings of the Association for Computational Linguistics: EMNLP 2023, 13171–13189. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Laverghetta, A., & Licato, J. (2023). *Generating better items for cognitive assessments using large language models* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/rqa9m

Lazarsfeld, P. F., Berelson, B., & Gaudet-Erskine, H. (1944). *The people's choice: How the voter makes up his mind in a presidential campaign* (Legacy edition). Columbia University Press.

Lebrun, B., Temtsin, S., Vonasch, A., & Bartneck, C. (2023). *Detecting the corruption of online questionnaires by artificial intelligence* (arXiv:2308.07499). arXiv. https://arxiv.org/abs/2308.07499

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, *38*(1), 163–190. https://doi.org/10.1007/s10869-022-09864-6

Li, B., Haider, S., & Callison-Burch, C. (2024). This land is your, my land: Evaluating geopolitical bias in language models through territorial disputes. In Proceedings of the 2024 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, June 2024, 3855–3871. https://doi.org/10.18653/v1/2024.naacl-long.213

Liew, A., & Mueller, K. (2022). *Using large language models to generate engaging captions for data visualizations* (arXiv:2212.14047). arXiv. https://arxiv.org/abs/2212.14047

Lipset, S. M., & Rokkan, S. (1967). Cleavage structures, party systems, and voter alignments. An introduction. In S. M. Lipset & S. Rokkan (Eds.), *Party systems and voter alignments: Cross-national perspectives* (pp. 1–64). Collier-Macmillan.

Lutz, C. (2019). Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies*, *1*(2), 141–148. https://doi.org/10.1002/hbe2.140

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of autoregression: Understanding large language models through the problem they are trained to solve* (arXiv:2309.13638). arXiv. https://arxiv.org/abs/2309.13638

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., & Miori, M. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*. https://doi.org/10.1177/20531680241231468

Mendoza, D. (2024, November 6). AI polling company defends wrong predictions on the US election. *Semafor*. https://www.semafor.com/article/11/06/2024/ai-startup-aaru-defends-using-artificial-intelligence-for-polling

Motoki, F., Neto, V. P., & Rodrigues, V. (2023). More human than human: Measuring ChatGPT political bias. *Public Choice*, *198*(3), 3–23. https://doi.org/10.1007/s11127-023-01097-2

Muhlberger, P. (2003). Political values, political attitudes, and attitude polarization in internet political discussion: Political transformation or politics as usual? *Communications*, *28*(2), 107–133. https://doi.org/10.1515/comm.2003.009

Nicoletti, L., & Bass, D. (2023, October 3). Humans are biased. Generative AI is even worse. *Bloomberg.Com*. https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Nie, E., Shao, B., Ding, Z., Wang, M., Schmid, H., & Schütze, H. (2024). *BMIKE-53: Investigating cross-lingual knowledge editing with in-context learning* (arXiv:2406.17764). arXiv. https://arxiv.org/abs/2406.17764

Nie, E., Yuan, S., Ma, B., Schmid, H., Färber, M., Kreuter, F., & Schütze, H. (2024). *Decomposed prompting: Unveiling multilingual linguistic structure knowledge in English-centric large language models* (arXiv:2402.18397). arXiv. https://arxiv.org/abs/2402.18397

OpenAI. (n.d.). *Do the OpenAI API models have knowledge of current events?* https://help.openai.com/en/articles/6639781-do-the-openai-api-models-have-knowledge-of-current-events

OpenAI. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., & Zoph, B. (2023). *GPT-4 technical report* (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2303.08774

OpenAI. (2024a, July 2). *GPT-4 API general availability and deprecation of older models in the Completions API*. https://openai.com/index/gpt-4-api-general-availability/

OpenAI. (2024b, July 2). *Models—OpenAI API*. https://platform.openai.com/docs/models/gpt-3-5-turbo

R Core Team. (2024). *R: The R project for statistical computing*. https://www.r-project.org/

Roth, M. (2024). *Ask a Llama - Creating variance in synthetic survey data*. General Online Research (GOR 24), Rheinische Fachhochschule Köln. February 22, 2024. https://doi.org/10.17605/OSF.IO/CHBTR

Rubin, D. B. (1987). *Multimple Imputation for Nonresponse in Surveys*. Wiley.

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023). *Towards coding social science datasets with language models* (arXiv:2306.02177). arXiv. https://arxiv.org/abs/2306.02177

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). *Whose opinions do language models reflect?* (arXiv:2303.17548). arXiv. https://arxiv.org/abs/2303.17548

Sass, K., & Kuhnle, S. (2023). The gender cleavage: Updating Rokkanian theory for the twenty-first century. *Social Politics: International Studies in Gender, State & Society*, *30*(1), 188–210. https://doi.org/10.1093/sp/jxac003

Schmitt-Beck, R., Roßteutscher, S., Schoen, H., Weßels, B., & Wolf, C. (Eds.), (2022a). *The changing German voter* (1st ed., pp. 313–336). Oxford University Press. https://doi.org/10.1093/oso/9780198847519.001.0001

Schmitt-Beck, R., Roßteutscher, S., Schoen, H., Weßels, B., & Wolf, C. (2022b). A new era of electoral instability. In R. Schmitt-Beck, S. Roßteutscher, H. Schoen, B. Weßels, & C. Wolf (Eds.), *The changing German voter* (1st ed., pp. 3–26). Oxford University PressOxford. https://doi.org/10.1093/oso/9780198847519.001.0001

Schoen, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., & Wolf, C. (Eds.), (2017). *Voters and voting in context: Multiple contexts and the heterogeneous German electorate* (1st ed.). Oxford University Press.

Schumacher, S., & Kent, N. (2020). *8 charts on internet use around the world as countries grapple with COVID-19*. Pew Research Center. https://www.pewresearch.org/short-reads/2020/04/02/8-charts-on-internet-use-around-the-world-as-countries-grapple-with-covid-19/

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, *616*(7957), 413. https://doi.org/10.1038/d41586-023-01295-4

Sturgis, P., & Luff, R. (2021). The demise of the survey? A research note on trends in the use of survey data in the social sciences, 1939 to 2015. *International Journal of Social Research Methodology*, *24*(6), 691–696. https://doi.org/10.1080/13645579.2020.1844896

Sultanum, N., & Srinivasan, A. (2023). *DataTales: Investigating the use of large language models for authoring data-driven articles* (arXiv:2308.04076). arXiv. https://arxiv.org/abs/2308.04076

Synthetic Users. (n.d.). *Synthetic users*. https://www.syntheticusers.com

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 (version 1)*. arXiv. https://doi.org/10.48550/ARXIV.2302.13971

Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3144139

Turk, V. (2023, October 10). *How AI reduces the world to stereotypes*. Rest of World. https://restofworld.org/2023/ai-image-stereotypes/

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in *R*. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed., corr. print). Springer.

Villalba, A. C., Brown, E. M., Scurrell, J. V., Entenmann, J., & Daepp, M. I. G. (2023). *Automated interviewer or augmented survey? Collecting social data with large language models* (arXiv: 2309.10187). arXiv. https://arxiv.org/abs/2309.10187

von der Heyde, L., Haensch, A.-C., Ma, B., & Wenz, A. (2025). *United in diversity? Contextual biases in LLM-based predictions of the 2024 European parliament elections* (arXiv: 2409.09045). arXiv. https://arxiv.org/abs/2409.09045

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J., Tu, Z., & Lyu, M. R. (2023). *Not all countries celebrate thanksgiving: On the cultural dominance in large language models* (No. arXiv:2310.12481). arXiv. https://arxiv.org/abs/2310.12481

Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., & Plank, B. (2024). *"My answer is C": First-token probabilities do not match text answers in instruction-tuned language models* (arXiv:2402.14499). arXiv. https://arxiv.org/abs/2402.14499

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wikipedia. (2024). Wikipedia:Wikipedians—Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedians&oldid=1184672006#cite_note-UNU-M-7

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, *38*(1–2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge University Press.

W3Techs. (2024). Usage statistics and market share of content languages for websites, July 2024. https://w3techs.com/technologies/overview/content_language

Author Biographies

Leah von der Heyde is a Researcher at the Social Data Science and AI Lab at LMU Munich and a Junior Fellow at the Munich Center for Machine Learning. Her research centers around the potentials and pitfalls of large language models, digital traces, and social media data for improving social science data quality.

Anna-Carolina Haensch is an Assistant Research Professor at the University of Maryland and a Researcher at LMU Munich. Her research focuses on multiple imputation, synthetic data, and large language models.

Alexander Wenz is a Research Fellow at the Mannheim Centre for European Social Research at the University of Mannheim. His research investigates the quality of novel methods of data collection, with a focus on mobile web surveys, smartphone apps, wearable sensors, and digital behavioral data.