Who Counts? Survey Data Quality in the Age of AI

Inaugural dissertation submitted in partial fulfillment of the requirements for the degree
Doctor of Social Sciences at the
University of Mannheim

by

Leah von der Heyde

Dean of the School of Social Sciences:
Dr. Julian Dierkes

First supervisor:
Dr. Alexander Wenz

Second supervisor:
Prof. Dr. Frauke Kreuter

Thesis reviewers:
Prof. Dr. Frauke Kreuter
Prof. Dr. Florian Keusch

Date of Defense:
July 1, 2025

Acknowledgments

This dissertation would not exist without the support of many people.

First of all, I am deeply grateful to my fantastic supervisor and co-author Alexander Wenz for his encouragement, feedback, and time. The harmony, trust, and mutual understanding that characterized this mentorship are rare and special, and made all the difference.

I also thank Caro Haensch for her continuous support, both as a mentor who always believed in me and taught me invaluable lessons in being an academic researcher, and as co-author of three papers of this dissertation.

Furthermore, I want to express my gratitude to Frauke Kreuter for inviting me to join the SODA group and funding me through the MCML, for getting me to think about data quality in new ways that were essential for this dissertation, and for challenging me to reach beyond what I thought was possible. I thank her and Florian Keusch for reviewing my dissertation, and for creating and leading a supportive community of stellar researchers.

Thank you to this very FK2RG research group for their rigorous feedback on my work and for being an inspiring community – sharing and critically discussing research, exchanging experiences, and sharpening skills. A special shoutout goes to my fellow PhD colleagues in Munich and Mannheim for the refreshing chats – especially Johanna Hölzl and Jacob Beck.

Further, I would like to thank Bernd Weiß, Jessica Daikeler, and Bolei Ma for their respective contributions as co-authors of papers featured in this dissertation, and Frederic Gerdon and Mariel Leonard for the insights and discussions while jointly working on other projects. Special thanks also go to Wiebke Weber not only for managing the administrative complexities to ensure we could smoothly carry out our research, but also for her guidance in many conversations beyond specific projects.

Last, but definitely not least, a huge Thank you to my loved ones for encouraging me to start this endeavor and for their emotional support throughout the ups and downs over the past years – to my friends in Sweden for the countless remote work sessions and fikas, and especially to my parents and friends in Munich for welcoming me back and supporting me in making it to the finish line.

Contents

1 Introduction

In the past decades, survey research has continually been facing challenges. Certain population segments, such as those with attitudes and behaviors likely different from the "average" person, or marginalized groups, are ever harder to reach under probability sampling frames (Lyberg et al., 2014). Response rates in traditional modes, such as face-to-face or telephone surveys, are declining (Luiten et al., 2020). Self-administered modes, such as mail or web surveys, can help obtain responses to sensitive questions and reduce social desirability, but also introduce new measurement errors (Tourangeau & Yan, 2007). This era of challenges, however, has coincided with the advancement of the digitalization of society, allowing survey methodologists to explore a variety of new modes and data sources for collecting or supplementing survey data. Accelerated by the necessities of the coronavirus pandemic, the use of the Internet for data collection in the form of (mobile) web surveys is now common (Gummer et al., 2023; Kennedy, Popky, & Keeter, 2023; Kohler, 2020). For example, it has become increasingly popular to sample and recruit survey participants through the Internet (e.g., on social media platforms), especially those belonging to specific subgroups underrepresented in traditional sampling frames. Additionally, there has been a proliferation of research evaluating and using digital trace data directly for estimating people's attitudes and behaviors (Conrad et al., 2021), including, among others, social media (Murphy et al., 2014), mobile app (Struminskaya et al., 2020), and Internet search data (Hölzl et al., 2025). Such data tends to be organic and observational – and not generated with the primary goal of producing population-level estimates and correlates (Groves & Lyberg, 2010; Salganik, 2019). As a result, the data comes with various challenges regarding representation and measurement. For example, coverage error might arise due to discrepancies between the target population and the user population of the specific data source (e.g., Hargittai, 2020). Validity issues might arise, for instance, because online behaviors are not always valid indicators of attitudes or behaviors (e.g., Bradley et al., 2021; Jungherr et al., 2017). Such errors jeopardize data quality – as a consequence, inferences made might not be accurate, ultimately leading to skewed or outright wrong understandings of society. Traditionally, total survey error frameworks (see, e.g., Groves et al., 2009) have been developed and used for identifying and quantifying the errors that can arise at different steps of the survey research process. The technological developments of the past decade have prompted the adaptation of these frameworks to new data sources, such as Big Data (Amaya et al., 2020), digital trace data (Sen et al., 2021), and more (see Daikeler et al., 2024, for a review), highlighting the conditions for using digital data for valid inferences about human attitudes and behavior.

More recently, as society moved into the age of AI, large language models (LLMs) entered the survey research chat. With their multi-purpose and multi-lingual generative capacities, these sophisticated machine learning models have been the target of large hopes for alleviating existing challenges in survey research. For example, a recent study on the impact of generative AI on the labor market estimates that 75-84% of survey researchers' tasks are exposed to LLMs, reducing time spent on those tasks by at least 50% (Eloundou et al., 2024). However, as they are based on Internet data, LLMs may come with similar potential pitfalls as other digital data sources

with regard to making inferences about human attitudes and behavior. As such, they not only have the potential to mitigate, but also to amplify existing biases regarding our understanding of different populations and constructs of interest. Due to selective corpora used for building LLMs as well as the digital divide (Lutz, 2019), *who is being counted* in LLMs' input and output data likely does not represent all populations and their subgroups equally, both in terms of scope and quality. In addition, the reliance on LLMs with their idiosyncratic data-generating processes puts into question *who does the counting* – researchers, data curators, data annotators, machines? – and *what is being counted and how it is being counted* – is the most likely next word in a sentence (what an LLM predicts) a valid indicator of social science concepts that have been tried and tested in surveys? Overall, *who counts* in survey research, even or especially in the age of AI, thus hinges on these fundamental questions of representation and measurement. In order to make valid inferences, it is important that the diversity of a population's perspectives is accurately measured and represented, both from a statistical and ethical point of view. Biased data of how people think and act could lead to misinformed and therefore ineffective or illegitimate policy decisions, which might unequally affect different subgroups. As a result, it could erode social cohesion and the trust in research and democracy (see, e.g., Nie, 2024). In short, LLMs have the potential to revolutionize survey and social science research (Bail, 2024; Grossmann et al., 2023; Ziems et al., 2024), with potentially serious consequences for society – for better or for worse.

The goal of this dissertation is to investigate whether and under which conditions LLMs can be leveraged in survey research by providing empirical evidence of the potentials and limits of their applications. Thereby, I want to inform the debate around whether LLMs are a general-purpose or specialized tool in the survey researcher's toolbox – can they be used for any survey-research related task or just for selected tasks? – and to contribute to the further improvement of LLMs for survey research and development of LLM-based survey methodology. I do this by applying LLMs to situations that go beyond so-called "high-resource" tasks. *High-resource* tasks are tasks that are easy for an LLM to complete because it has been provided with more training data enabling it to fulfill the tasks – for example, common logical problems (McCoy et al., 2023) or English-language text (Dey et al., 2024). This stands in contrast to "low-resource" tasks, for example, niche applications or languages. More specifically, I test two major potential applications of LLMs in survey research – simulating respondents and coding open-ended responses – in previously unexamined contexts – European societies and languages.

The remainder of this introduction is structured as follows. First, I introduce LLMs and key related concepts used in this dissertation, and provide an overview of the potential applications of LLMs at different stages in the survey research process. I then discuss the potential challenges of such applications and their consequences for data quality. I argue that, due to these challenges, the proposed applications need to be investigated systematically, which presents an important research gap. Next, I highlight the contributions of this dissertation in relation to this gap. Finally, I provide a summary of the following chapters of this dissertation, which feature research rigorously testing the aforementioned applications of LLMs regarding the representation and measurement of human attitudes and behavior.

1.1 A Small Introduction to Large Language Models

High-quality data about human attitudes and behavior lies at the heart of answering many social science research questions, and surveys are one of the most popular tools to obtain such data (Couper, 2013; Grossmann et al., 2023; B. J. Jansen et al., 2023). To assess how LLMs might help

or hinder in this effort, it is necessary to understand how they work. LLMs like GPT (OpenAI et al., 2023), Llama (Dubey et al., 2024), Claude (Anthropic, 2025), or DeepSeek (DeepSeek-AI et al., 2025) are a form of generative artificial intelligence, designed to process and generate human-intelligible text across a wide range of topics. Multimodal models can also process and/or generate image (e.g., Dall-E, Rombach et al., 2022, Stable Diffusion, Ramesh et al., 2021), audio (e.g., Whisper, Radford et al., 2022), and video (e.g., Sora, OpenAI, 2024) data. LLMs are trained on large amounts of Internet text data, such as selected book collections, Wikipedia entries, and social media data (e.g., Brown et al., 2020; Roberts, 2022). Dataset-based training (i.e., linguistic learning and "knowledge" building) is complemented by training for following instructions (i.e., preparing for natural language user input) as well as human feedback on responses (i.e., optimizing for desired output). This way, LLMs learn to internalize patterns, structures, and contextual relationships between words in human-generated texts, and to predict missing or next words in a sequence. Given a user input in natural, i.e., human, language (*prompt*), they convert the text into numerical representations (*tokens*), analyze the relationships between these tokens based on their learned knowledge from the training data, and then, conditional on the previous input and output words, iteratively predict the most likely next token. The result is textual output (a *completion*) in the form of natural language, making them *generative* LLMs.

For the purpose of this dissertation, the term "LLMs" refers to such generative LLMs. In contrast, language models of the BERT family, which have become popular tools for a range of text-related computational social science tasks (see, e.g., Wankmüller, 2024, for a review), are *analytical* LLMs, designed primarily for understanding and classifying text rather than generating it. While BERT is trained to fill in missing words and predict relationships between sentences, it needs a specific context (e.g., a full sentence) to complete such tasks and needs to be fine-tuned to be effective at tasks like classification, sentiment analysis, and named-entity recognition. *Fine-tuning* involves further training of language models with a dataset of input-output pairs for the specific use case. However, BERT does not handle incomplete (e.g., "I will vote for . . . ") or open-ended prompts (e.g., "Describe people who support the Conservative Party") well.

Unlike analytical LLMs, generative LLMs are capable of understanding and generating text dynamically, enabling them to create fluent and coherent responses rather than just analyzing or classifying text. As general-purpose LLMs, they are able to complete a broad range of text-processing tasks without necessarily needing fine-tuning. Their usability with *few-shot* prompting – i.e., including a handful of examples of input and desired output directly in the prompt – or even *zero-shot* prompting without such examples, reduces the need for several different task-specific (analytical) LLMs. Although usually optimized for English, generative LLMs are generally multilingual. These features enable them to summarize information, translate languages, answer complex questions, write computer code, and engage in nuanced human-like conversations in multiple languages. Their capacity for processing and generating natural language and availability as both chat-based interfaces (e.g., ChatGPT) and through Application Programming Interfaces (APIs) makes them an easily accessible tool for (computational) social scientists. These features render LLMs particularly relevant for applications in social science research more generally and in survey research in particular, where they are likely to become a standard tool (Bail, 2024; Demszky et al., 2023; Grossmann et al., 2023; Ziems et al., 2024). In the existing literature, three main potential application areas for LLMs within the survey research process have emerged (Bail, 2024; Kreuter, 2025) – LLMs acting as interviewers, LLMs acting as respondents, and LLMs acting as research assistants. However, use cases for LLMs are conceivable in virtually all stages of the survey research process, where they could possibly help address errors impacting data quality

(Barari et al., 2024). The following section provides an overview of the potential applications of LLMs in the survey research process before, during, and after data collection.

1.2 Potential Applications of LLMs in the Survey Research Process[1]

Pre-data collection

Built for creative text generation, LLMs could assist in the **questionnaire design** phase by developing new questions (Götz et al., 2023; Hernandez & Nie, 2022; Konstantis et al., 2023; Laverghetta Jr. & Licato, 2023; Lee et al., 2023; Maiorino et al., 2023; Zou et al., 2024), items for scales and indices (Sarstedt et al., 2024), experimental vignettes or images (Bail, 2024; Demszky et al., 2023; Sarstedt et al., 2024), or entire questionnaires. They can also evaluate existing questions (Hommel, 2023; Olivos & Liu, 2024), including assessing their readability or social desirability, detecting leading or biased wording, simplifying or adjusting language for different literacy levels and cultural contexts, and suggest concrete improvements (Jacobsen et al., 2025; Thirunavukarasu & O'Logbon, 2024). Another use case for LLMs in this stage is translating questionnaires, either by providing multilingual translations with context-aware adjustments (Adhikari et al., 2025), or by checking existing translations for accuracy, consistency, and meaning preservation. As such, LLMs could be integrated as one of the translators or as an adjudicator in the TRAPD approach (Translation, Review, Adjudication, Pretest, Documentation; Harkness, 2003), which usually features two independent human translations and a human adjudicator in case of disagreements. These applications in questionnaire design have the potential to facilitate the increase of validity and reduction of measurement error, especially the one inadvertently introduced by human researchers.

During **pre-testing**, LLMs have the potential to mitigate measurement error by analyzing responses from pilot surveys, identifying patterns, and suggesting modifications (Kreuter, 2025). Taking the integration of LLMs into pre-testing processes even further, they could act as virtual or simulated respondents (see subsection 1.2.2 for a more detailed explanation of so-called "synthetic samples"). In the form of audio or video avatars, such virtual respondents could be used for **interviewer training** ahead of personal interviews (Thirunavukarasu & O'Logbon, 2024), simulating diverse groups' interpretations of and reactions to the questions (Dillion et al., 2023; Grossmann et al., 2023).

Regarding **sampling** and **recruitment**, LLMs could aid in defining the target population based on the research questions and analyses of previous surveys and research papers, suggesting appropriate sampling frames. They could summarize best practices and recommend different types of sampling designs. Further, they could also review sampling plans, highlight potential biases or limitations, and make suggestions for improvement. They could also possibly reduce sampling error more practically (Barari et al., 2024), for example in the processing of address-based samples. Closer to their original purpose of creatively generating human-like text, LLMs could be used for creating recruitment material and adapting it to different outreach formats, such as mail, e-mail, social media advertisements, or verbal recruitment scripts, which could aid in reducing nonresponse error based on unit nonresponse. While LLMs thus could potentially be used in these parts of the survey research process, this has not been done in prior research.

[1] This section, together with section 1.3, has been published as a conference paper at COLM as von der Heyde, L. (2025). Who Counts? The Potentials and Pitfalls of Using LLMs in Survey Research. *First Workshop on Bridging NLP and Public Opinion Research.* https://openreview.net/forum?id=ww2KqnPLdK

Data Collection

LLMs have major potential in the data collection phase of survey research. Here, they could augment surveys by **dynamically adapting the questionnaire** based on previously given responses, for example through probing questions (Barari et al., 2024; Geisen, 2024) or by devising and deploying real-time strategies for reducing respondent burden, inattentiveness, item nonresponse, or breakoff. Dynamic, LLM-generated probing questions could also be used to scale up in-depth interviewing by integrating them into web surveys (Jacobsen et al., 2025). Another example is the automatic creation and real-time fielding of new survey items from open-ended responses to allow for the standardized measurement of emerging relevant topics within specific populations (Velez, 2025). The increased relevance and responsiveness of such dynamic surveys could help improve survey engagement and reduce breakoff, thereby reducing both measurement and nonresponse error.

Beyond assisting human interviewers, LLMs could also be deployed as **independent interviewers** conducting text- or voice-based interviews (Barari et al., 2024; Grossmann et al., 2023; Lerner, 2024), allowing for the implementation of automated conversational interviewing. In *web-based* surveys, for example, LLMs can be set up as chatbots for creating an online text-based conversational interviewing format for self-administration (Cuevas et al., 2023; Wuttke et al., 2024; Xiao et al., 2020; Zarouali et al., 2023). Alternatively, they can power artificial audio or video avatars (akin those offered by, e.g., Tavus, 2025) in more traditional web survey formats. *Telephone surveys* with LLM-based interviewers (e.g., Lang & Eskenazi, 2025) present another option. Regardless of mode, these implementations of dynamically responsive LLM interviewers would make such semi-automated, "personal" survey administration more flexible than the pre-programmed versions of previous decades (Conrad et al., 2015, 2019). This way, LLMs might be able to help address comprehension issues by providing examples or answering respondents' follow-up questions that may not be accounted for in standardized web surveys or interview protocols (B. J. Jansen et al., 2023), and possibly ease participation by visually impaired persons.

However, the effect of these innovations on data quality is still unclear. While augmenting or replacing human interviewers could have a positive impact on response quality and completion rates, thereby reducing measurement and nonresponse error (Lerner, 2024), the lack of human touch could also lead to less engagement, acting in the opposite direction (Lang & Eskenazi, 2025).

Another prominent application of LLMs is that of simulating respondents through LLM-based **synthetic samples**[2], which can be relevant for several stages of the survey research process. To create them, an LLM is prompted to generate an artificial dataset of survey responses to the question(s) of interest, which can then be used in the context of pre-testing, simulation analyses, primary data collection, or imputation. In their most basic form, the LLM could repeatedly be asked to respond to survey questions. In order to better approximate specific target populations and their response distributions, such samples can also be created based on "personas", by sequentially feeding individual information about humans, for example socio-demographic and attitudinal information collected in surveys, to an LLM, and asking it to respond to survey questions

[2] "Synthetic samples" should be distinguished from the "synthetic data" that is used in the context of more traditional imputation and anonymization processes. While LLM-synthetic samples could be used for such purposes, the underlying statistical assumptions and calculations are more advanced in the traditional sense of the term. Other terms used in the literature on LLM-synthetic samples are "silicon samples" or, when specifically mirroring specific (types of) respondents, "personas" or "subpopulation representative models". These terms are used interchangeably in this dissertation.

from the respective person's perspective (e.g., Argyle et al., 2023; Bisbee et al., 2024; Simmons & Hare, 2023). This can be done either by providing answer options and asking for a verbatim or option number/letter response, or by requesting an "open-end" response that maps onto the closed-ended question. The persona-based approach allows researchers to simulate a vast array of individual positionalities and perspectives, which has been argued to address generalizability concerns (Grossmann et al., 2023) and help address coverage and sampling errors. Because of this, some researchers argue that LLM-based synthetic samples are better-suited for social science research than the convenience samples used in many studies (Bail, 2024). Depending on the amount of questions asked, both univariate and joint distributions could be modeled based on such samples (Simmons & Hare, 2023).

Synthetic samples could be used to supplement existing survey data, by imputing missing data due to unit- or item-nonresponse, for example on sensitive topics or with hard-to-reach populations (Grossmann et al., 2023; B. J. Jansen et al., 2023; Kalinin, 2023) or by generating data for single items previously unasked (J. Kim & Lee, 2023). Other potential advantages proposed are that synthetic respondents do not require the creation of complex sampling schemes or costly incentives (Dillion et al., 2023), and might not exhibit human response bias or interview fatigue (e.g., Dillion et al., 2023; Grossmann et al., 2023; B. J. Jansen et al., 2023, but see subsection 1.3.2). Some researchers have even suggested LLM-synthetic samples could completely substitute survey data (e.g., Aher et al., 2023; Argyle et al., 2023; Horton, 2023; see Agnew et al., 2024 for a review of positions). Furthermore, such samples could be employed for pre-testing surveys (e.g., Webb, 2024), thus saving resources needed for actual surveys of humans for the main fieldwork (e.g., Hewitt et al., 2024). For example, they can be used for conducting preliminary analyses (Bail, 2024; Sarstedt et al., 2024; Thirunavukarasu & O'Logbon, 2024), allowing for, e.g., estimation of effect sizes for hypothesis generation, or power analyses for optimal sample design (Demszky et al., 2023; Grossmann et al., 2023). Finally, (partially) substituting human participants with LLM-generated counterparts could reduce respondent burden, for example by minimizing harm in case of potentially distressing or sensitive survey topics or experiments containing misinformation, or simply by reducing the amount of questions respondents need to be asked. As such, the use of LLM-synthetic samples can be situated either at the step of instrument design, data collection, or processing in the survey life cycle, and could potentially aid in reducing four major components of total survey error – coverage, sampling, measurement, and nonresponse.

Post-data collection

Upon the completion of data collection, multimodal LLMs could help with **data processing** by digitizing survey data, for example by transcribing audio data from in-person, phone, or web-based interviews (Revilla et al., 2025; Tewari & Hosein, 2024). Transforming scans of paper-based (mail) questionnaires into tabular data with optical character recognition is another (yet to be explored) possibility, which would render specialized machines for such efforts obsolete. LLMs could also aid in structuring previously unstructured data used to augment surveys for learning about attitudes and behaviors, such as social media data (Cerina & Duch, 2023) or data donations, i.e., individual-level digital behavioral data, such as mobile usage or social media data donated by users themselves (Carrière et al., 2024). More generally, they could perform a range of code-based data wrangling tasks (Jaimovitch-López et al., 2023). Such applications could reduce human-generated processing errors.

LLMs could then be used for quality checks, further mitigating measurement error. They

could detect low-quality or outright fraudulent responses by analyzing response patterns and identifying inconsistent responses based on time taken to complete the survey, contradictory statements in scales, or the content of open-ends (Lebrun et al., 2024). This applies not just to human responses: Because LLMs can not only be used for detecting fraudulent responses, but also for creating them (Veselovsky et al., 2025), detecting such LLM-bot responses can be achieved through prompt injections in questionnaires targeting LLMs (Höhne et al., 2025) or even be aided by LLMs (Lerner, 2024). For personal interviews, LLMs could also check interviewer adherence to the interview scripts by matching them against the interview transcripts, safeguarding measurement quality.

Further, LLMs could aid in data processing by coding[3] text, image, or audio data (see Ziems et al., 2024, for a systematic review of using LLMs for coding social science text). In the survey context, such data can, for example, come from open-ended responses, social media, or surveys asking respondents to upload pictures of their surroundings (Bail, 2024; Demszky et al., 2023; see Iglesias et al., 2024 for an illustration). The advantage of using LLMs lies in their speed and scalability, allowing researchers to code an entire corpus of data instead of just a sample. Examples of such applications include sentiment analysis, named-entity recognition, identifying political affiliations, or the presence or absence of a specific concept (Ahnert et al., 2025; Bail, 2024; Cerina & Duch, 2023; Demszky et al., 2023; Gilardi et al., 2023; Törnberg, 2024). Beyond such coding tasks with a predefined coding scheme, LLMs could be asked to develop coding schemes based on theory or based on the data given, i.e., unsupervised labeling or topic modeling (Ornstein et al., 2024; Pham et al., 2024). Researchers hope that LLMs could minimize human coders' subjectivity, inconsistency, and lack of attention (Bail, 2024), thereby minimizing measurement and processing error – however, human validation is still recommended (see also Chapter 4).

Also addressing processing error, LLMs could generate standardized and easy-to-use variable labels for datasets. Given information about the data structure, they can assist in writing code for data processing and analysis in a variety of programming languages, such as R or Python. Finally, LLMs could assist in calculating and adjusting survey weights based on census data. With harmonizing efforts, LLMs could furthermore efficiently match and map variables from different surveys to ensure comparability, or even help integrate social media or administrative data and survey data (B. J. Jansen et al., 2023).

During **data analysis**, LLMs could assist by summarizing tabular (quantitative), textual (open-ended or qualitative interview), or audio survey data into text, providing both high-level overviews and detailed findings (Thirunavukarasu & O'Logbon, 2024). They could also be used for (writing code for) generating data visualizations or for generating captions for existing ones (Liew & Mueller, 2022; Thirunavukarasu & O'Logbon, 2024; C. Wang et al., 2025). Ultimately, LLMs could draft complete reports based on structured survey data (Sultanum & Srinivasan, 2023).

As is evident from this review of potential applications, LLMs could help make survey research more efficient, while also reducing some common forms of human-induced errors. After all, the overall aim of survey methodology is optimizing data quality. However, as has been the case for other new forms of data, methods, and technology (e.g., Couper, 2013; Sen et al., 2021), integrating LLMs into the survey research process can also incur new forms of errors and issues that survey methodologists need to be aware of. The following section points out some of the potential challenges for data quality when using LLMs in survey research.

[3]The terms "coding", "classifying", and "labeling" are used interchangeably in this dissertation.

1.3 Data Quality Challenges in LLM-Based Survey Research

Although LLMs have only been widely discussed in research and society relatively recently, biases in their outputs were quickly identified. These biases relate to aspects of central importance for social science research: LLMs exhibit general cultural biases, including a tendency towards reflecting or assuming Western or U.S.-centric norms, idealizing whiteness and masculinity, and an inability to replicate other cultural values (e.g., Atari et al., 2023; Bianchi et al., 2023; Havaldar et al., 2023; Johnson et al., 2022; Masoud et al., 2025; Palacios Barea et al., 2023; Ramezani & Xu, 2023). Also when it comes to psychological measures, LLMs have been shown to be WEIRD – they mostly resemble Western, Educated, Industrialized, Rich, and Democratic populations (e.g., Atari et al., 2023; but see Niszczota et al., 2025). Politically, several studies suggest that the default outputs of LLMs skew left (e.g., Batzner et al., 2024; Hartmann et al., 2023; Motoki et al., 2023; Rettenberger et al., 2025), partially moderated by the assumed ideology of populations using the input language (Li et al., 2024; Walker & Timoneda, n.d.). Further, LLMs exhibit worse performance in non-English languages (e.g., Schott et al., 2023; Zhang et al., 2023), reproducing assumptions and stereotypes associated with English-speaking contexts (Ghosh & Caliskan, 2023; Öztürk et al., 2025; W. Wang et al., 2024). Even in English, LLMs reproduce negative stereotypes about sexual and racial minorities and more complex intersectional identities (Gross, 2023; Gupta et al., 2024; Hada et al., 2023; Haim et al., 2024; Ma et al., 2023; Nagireddy et al., 2024; Ostrow & Lopez, 2025).

Such biases in LLM outputs can stem from multiple underlying roots (Hovy & Prabhumoye, 2021; McCoy et al., 2023). These include the pre-determined input provided to LLMs, i.e., training data, annotation, and alignment processes; the model architecture, i.e., their purpose and design; and the research design, i.e., prompting and hyperparameters controlled by the researchers themselves. Biases in these roots can have direct impacts on the quality of survey data generated, processed, and analyzed with the help of LLMs.

Training and alignment

When it comes to LLM training data, it is important to note that these training corpora contain a large, but not balanced selection of human-generated text. The corpora likely[4] do not feature the diversity of attitudes and behaviors present in human populations, due to a dual selection bias: the digital divide impacts the composition of the "sampling frame" of potential training texts representing humans vis-à-vis the target populations. The non-randomness of texts selected for training corpora impacts the composition of the "sample" of actual training texts vis-à-vis the "sampling frame".

Regarding the **digital divide**, bias is potentially introduced at several levels: First, one must consider that there are *cross-national* differences in Internet access and behavior (International Telecommunication Union, 2022; Schumacher & Kent, 2020). Although global Internet penetration rates are by now high, people without Internet access almost exclusively live in non-WEIRD countries (Crockett & Messeri, 2023; International Telecommunication Union, 2022). Second, there are *cross-sectional* differences related to platform selection, production of Internet text, and type of text production. These differences include sociodemographic, socioeconomic, and attitudinal factors, such as age, education, and ideology (Blank, 2013; Hargittai, 2020; Hoffmann

[4] The opacity of LLM training data makes the identification of underlying biases challenging and speculative (e.g., Bail, 2024; Kuntz & Silva, 2023).

et al., 2015; J. W. Kim et al., 2021; Shaw & Hargittai, 2018; Tucker et al., 2018) and interact with the cross-national differences (Schumacher & Kent, 2020). Because of these disparities, even if Internet text was randomly selected for LLM training, certain populations and subgroups would be systematically under- or overrepresented. The determinants of differences in online behavior correlate with many key outcomes of interest in social science research (Dutwin & Buskirk, 2023). This can lead to coverage bias when using LLMs for survey research, as the attitudes and behaviors of, e.g., older, less educated or skilled people and such with marginalized identities are less likely to be featured in LLM input (and therefore, output), simply because they are featured less on the Internet (Crockett & Messeri, 2023). As a result, LLMs might struggle with accurately representing such groups or individuals when tasked to mimic respondents, code and analyze responses, or during questionnaire design and evaluation. For example, research suggests that LLMs are better able to emulate the attitudes of Western, English-speaking, developed populations, particularly the U.S. (Qu & Wang, 2024), and do not represent all demographic subgroups equally well, even within the U.S. (Bisbee et al., 2024; Sanders et al., 2023; Santurkar et al., 2023). Beyond such biases undermining the multivariate analyses social scientists typically care about, the lack of variance in responses observed in these and similar studies also raises questions about the feasibility of synthetic samples in pre-testing. For example, when conducting power analyses, LLM-generated data would suggest implausibly low sample sizes.

However, the **selection of LLM training data** is not random (see Clemmensen & Kjærsgaard, 2023, for a discussion of the distinction between representative vs. diverse data in AI). On the contrary, LLM training corpora tend to be composed of sources authored by rather homogeneous communities, such as curated books, the English Wikipedia, and Reddit (Brown et al., 2020; Kuntz & Silva, 2023; Roberts, 2022; Shaw & Hargittai, 2018). What is true for sampling in general and has been confirmed in the context of survey research using Big Data also holds for LLM training data and inferences based on them: bigger is not always better, as coverage bias can persist and might only be amplified (Bradley et al., 2021; Hargittai, 2015). Web scraping, which is used to create a large part of LLM training datasets, can lead to sampling bias (Foerderer, 2023). As a result, minority languages and the perspectives of certain (sub)populations are likely underrepresented (Buschek & Thorp, n.d. Kuntz & Silva, 2023), and the explicit and implicit attitudinal and behavioral biases expressed by the authors of the texts in the selected datasets not only get encoded, but disproportionately amplified in LLMs (Bender et al., 2021). For example, Heseltine and Clemm von Hohenberg (2024) found that LLMs performed worse when labeling non-English political texts. This could lead to a distorted image of how underrepresented groups think and act, based on generalization or (out-group) stereotypes, either explicit or implicit in the training data, rather than (in-group) authentic content (see also Demszky et al., 2023; Linegar et al., 2023). This has major implications for the data quality of synthetic samples generated with LLMs: they can only be as diverse as the populations on which they were trained (Dillion et al., 2023; Grossmann et al., 2023). This limitation can undermine the goals of supplementing traditional survey data for marginalized subgroups that are harder to survey – they likely cannot be captured by LLM-generated data either.

Measurement challenges also arise when considering that some of the data featured in LLM training corpora is not necessarily an objective reflection of human preferences. Social media users' interaction with platforms is a function of their affordances and algorithms. Individuals might use certain expressions to make their content more engaging (Buschek & Thorp, n.d.), i.e., findable, likeable, and shareable, leading to an overestimation of certain concepts. The fact that **digital behavioral data is not primarily generated for social science** data collection also introduces validity issues that transfer to LLMs, which during their training process might

infer concepts from this data that are not actually correct. For example, it has been shown that mentions of political content in social media are an indicator of attention to politics rather than support of the mentioned person or issue (Jungherr et al., 2017).

In addition to these potential biases associated with the training data, label bias can occur when considering the **attributes of the workers annotating** LLM training data and aligning LLMs through their feedback (Grossmann et al., 2023; Hovy & Prabhumoye, 2021). For example, intra- and interpersonal variance in motivation and attention during such tasks can lead to skews in the data LLMs learn from. More consequentially, systematic misinterpretations due to different backgrounds can occur. These include differing interpretations of constructs between annotators, as well as mismatches between a text's author's intended meaning and the annotator's interpretation, possibly due to linguistic or cultural unfamiliarity (c.f. D'Ignazio & Klein, 2020). While the former may lead to certain interpretations being overrepresented in LLMs or LLMs having no clear understanding of a concept when they should have, the latter can incur misreporting of human attitudes and behavior when using LLMs in survey research, both in terms of measurement and representation.

Finally, the **temporality** of training data implies that off-the-shelf LLMs are not by default up to date with current developments, including changes in language use and global political, economic, and social realities. This can lead to measurement and representational challenges when using LLMs in survey research, as LLMs may produce output based outdated understandings of attitudes and behaviors. For example, an LLM might wrongly label a survey response as not containing racist attitudes although the connotation of the term used has since changed to express racism (or vice versa, in the case of groups actively re-claiming previously derogatory terms, making them no longer racist), resulting in faulty measurement. Representational issues could arise if, e.g., training data cutoffs preclude the realignment of political ideology and attitudes. For instance, in the context of war, left-leaners have traditionally been considered more dove-ish, and right-leaners more hawkish, but this relationship has reversed in the context of the war in Ukraine – something LLMs fail to pick up on if their training data cut off before the invasion (Sanders et al., 2023).

Model architecture

LLMs' design and purposes can also impact output data quality for survey research. Off-the-shelf LLMs' **optimization processes** tend to focus on tasks and benchmarks that are not directly related to survey research applications (Huckle & Williams, 2025; Sarstedt et al., 2024). McCoy et al. (2023) demonstrate that LLM output is skewed towards tasks and problems that are known to be more commonly mentioned in Internet text, regardless of task complexity. This is likely also the case for survey research tasks in general (e.g., solving math problems as a "high-resource" task vs. simulating respondents as a "low-resource" task), and specific subtasks (e.g., simulating respondents of populations better represented through the training process as a higher-resource task vs. simulating underrepresented respondents). Relatedly, although LLMs have an extensive general vocabulary, they might have trouble with less common or domain-specific terms (B. J. Jansen et al., 2023). Thus, LLMs might only be useful for survey research in very constrained settings, for specific tasks, topics, and populations (Dillion et al., 2023). Accordingly, the majority of studies employing LLMs for survey-related tasks can be considered a lower bound (Bail, 2024), since they tend to focus on high-resource contexts: English-speaking and Western, predominantly U.S.-American populations (e.g., Argyle et al., 2023; Bisbee et al., 2024; Cerina & Duch, 2023; J.

Kim & Lee, 2023; Mellon et al., 2024; Rytting et al., 2023; Sanders et al., 2023; Santurkar et al., 2023). The representational and measurement issues detailed in subsection 1.3.1 might inhibit the generalizability of these studies' findings, and survey methodologists need to investigate whether LLMs, or a specific LLM (see subsection 1.3.3), is fit for the purpose they want to employ it for.

Further, how LLMs transform textual input into semantic representations can lead to biases. The associations LLMs generate between words during their training processes (**embeddings**) might be biased as a result of explicit or implicit biases in the training data. In addition to the biases listed above, such erroneous associations could arise from spurious correlations in the training data which the LLM identifies as a pattern, since it relies on the input as a representation of reality (Grossmann et al., 2023). Such biases might only be masked by debiasing efforts, i.e., the LLM is prevented from explicitly generating harmfully stereotypical output, but the underlying biases might still carry through the way it performs, e.g., labeling tasks.

In addition, measurement challenges arise when considering **what LLM output technically represents**: the conditional probability of the previous (prompt and completion) words being followed by said output. In other words, while LLMs produce human-like text output, it is unclear whether that output represents (and therefore can approximate) human cognitive processes (Dillion et al., 2023). This puts into question construct validity, as such probabilities might not actually reflect social science constructs, but statistical and semantic probabilities. More fundamentally, it is not entirely transparent how LLMs arrive at their ultimate output, also considering it is probabilistic rather than deterministic (e.g., Grossmann et al., 2023). However, understanding the data-generating processes behind social science data lies at the foundation of inference. Measurement quality is further complicated because the generated natural language outputs sometimes do not match what the underlying probabilities would suggest (X. Wang et al., 2024). Thus, whether researchers use the text output at face value or whether they work with the underlying probabilities makes a difference for inference.

Another model design aspect potentially leading to errors is LLMs' more **explicit purposes**. On the one hand, they tend to be programmed to always be helpful and provide a satisfactory and confident response – even when the information in the training data would suggest an ambiguous response or none at all, e.g., due to lacking information. While this may solve missing data problems commonly found in survey research, it does not mirror human reality. For example, when using LLMs to simulate respondents, LLMs might respond where (certain) humans would refuse. Although this feature is often presented as desirable or even the point of using LLM-generated data in the first place, it challenges the validity of LLM-generated responses, as they do not mirror human behavior. On the other hand, guardrails designed for ensuring LLMs do not give overly sexist, racist, or otherwise harmful responses could lead to such perspectives, which do exist among humans, not being captured by LLM output (Demszky et al., 2023; Grossmann et al., 2023). Similar to social desirability in surveys, this "machine desirability" can negatively impact measurement. Relatedly, due to their programmed goal of agreeableness, LLMs might have a tendency for acquiescence bias (Bail, 2024; Dentella et al., 2023).

Research design

Moving from developer-determined specifications to researcher-determined factors, data quality can also be impacted by the specific **choice of LLM**. Each LLM is made up of a unique combination of training data, alignment processes, weights, and overall model architecture. For example, it has been found that GPT base models, i.e., LLMs that have not undergone alignment

based on human feedback, tend to reflect more lower-income, conservative views, whereas instruction-tuned GPT models have a liberal elite bias (Dillion et al., 2023). Therefore, how accurately human attitudes and behaviors are represented in LLM-based survey research depends on the chosen LLM. The choice of LLM might in turn be impacted by its affordances, such as the accessibility, user interface, or usage limits. LLMs vary in speed and cost as well as optimization for specific languages or tasks. They might also have different default values for hyperparameters, which researchers might be induced to carry forward as to not "artificially" alter the model, possibly resulting in less-than-optimal and incomparable output. Therefore, different LLMs may perform differently given the same survey research process task – the question then is not only whether an LLM can perform a task, but which LLM. This poses a challenge for generalizability claims of which tasks can be augmented by LLMs, and for best practice recommendations. This challenge is compounded by the fast-paced (and often intransparent or uncontrollable) updates to LLMs, which may not always carry performance improvements for the specific task at hand, impacting reliability.

Further, the variability of **model hyperparameters** (i.e., temperature, sampling range, and repetition penalties) potentially inhibits data quality of LLM-based survey research. For example, while the amount of randomness in LLM outputs can be reduced by lowering the temperature hyperparameter, thereby increasing reliability by forcing the LLM to always pick the most likely option, this also reduces within-group variability to a level unlikely found in humans: If given two output choices, for example, two response options to an attitudinal question or two categories for a text classification, one with a probability of 0.51 and one with 0.49, an LLM with minimum temperature would be forced to always choose the option with 0.51, even though, in reality, almost 50% of cases fall into the other category. Although experiments with different hyperparameters can yield insights into their optimization, the exact impact of these variables on the data-generating process within LLMs is opaque, challenging validity.

Another challenge for data quality in LLM-based survey research is the sensitivity of LLMs to **prompt wording** (e.g., Bisbee et al., 2024; Gui & Toubia, 2023; Pezeshkpour & Hruschka, 2024). That is, the choice and order of words and response options in the prompt input can impact the output. While the survey pre-testing literature is informative about which subtle questionnaire changes induce changes in human response behavior (e.g., Schuman & Presser, 1996), there is no generalizable or systematic information about this for LLMs. For example, Tjuatja et al. (2024) found that LLMs do not mirror human response biases, but exhibit idiosyncratic ones. There is competing evidence regarding LLM robustness to the order of options in closed-ended questions (e.g., Moore et al., 2024 vs. Pezeshkpour and Hruschka, 2024). In addition, simply adding more information (e.g., more detailed category descriptions for coding open-ended responses, or more information about respondents to be impersonated) might not necessarily lead to better output quality; research indicates that LLMs do not retain all information equally well in longer prompts, but sometimes tend to "forget" the middle part (Liu et al., 2024). Whether "system" prompts specifying overall task context and behavior ahead of individual requests (e.g., "You are a thorough survey researcher") can improve this is subject to debate (e.g., Zheng et al., 2024 vs. Fröhling et al., 2024).

As this section has demonstrated, there are numerous potential sources of error and bias in LLMs, but it is not clear how exactly these errors and biases play out in survey research applications. LLMs have the potential to mitigate existing data quality challenges in survey research, increase them, or introduce new ones, with both representational and measurement-related consequences. Estimates derived from biased LLMs ultimately risk leading to wrong conclusions,

which have the potential to perpetuate existing stereotypes and inequalities in research and society (Bail, 2024; Lutz, 2019, c.f. Robinson et al., 2015; Shaw and Hargittai, 2018), or to result in misinformed decisionmaking in businesses and public policy (Hargittai, 2020; Sarstedt et al., 2024). Further, the potential errors and biases identified in this section put into question the generalizability of singular studies showcasing the successful application of LLMs to different survey research tasks. In fact, many of these studies have been conducted in high-resource contexts, i.e., on U.S.-based or English-language texts that likely are overrepresented in LLMs' training. Based on these considerations, I argue that any survey-related application of LLMs needs to be evaluated for the specific context it is to be employed in.[5] This is especially true when bearing in mind that LLMs were not a priori designed for survey research. Compared to the long history of survey research methods, LLMs have emerged rather recently, which is why their applications to the field and their potential challenges have yet to be systematically evaluated and addressed. The following section highlights the contributions this dissertation makes in addressing this gap.

1.4 Contributions of This Dissertation

The aforementioned potential challenges and their implications point to the need for systematic methodological research investigating biases in LLM-based survey research. Such research is not only important for survey methodologists and practitioners in helping them know the potentials and limits of this tool for ensuring high-quality data, but can also inform the development of LLMs, both specifically for survey research applications and more generally for mitigating their inherent biases. Since LLMs are a newly emerging tool in the survey research toolbox, filling this gap will require extensive and ongoing work in a fast-moving environment. This dissertation contributes to this effort by (1) *examining the potential and pitfalls of LLMs in two of the major applications of LLMs in the survey research process* – simulating respondents and coding text data – *in previously unexamined population contexts* – European countries and languages, and by (2) *widening the multinational, multilingual, and multicultural scope of these applications* – providing insights into the generalizability of their applicability beyond the initial high-resource population contexts (U.S.-based and English data) and beyond the specific test cases discussed here. One of the applications faces representational challenges – supplementing or substituting survey data with LLM-generated data (Chapters 2 and 3) – and the other measurement challenges – coding open-ended survey responses with LLMs (Chapter 4).

For both applications, and for LLM-based survey research in general, existing research has mostly been focusing on relatively "easy" tasks for LLMs – tasks that are likely completed successfully due to the high prevalence of relevant data in LLMs' training data (e.g., Argyle et al., 2023; Mellon et al., 2024). As elaborated above, this is the case for English-language tasks and tasks related to the attitudes and behaviors of the U.S. population. However, as cross-cultural survey researchers are aware of from traditional survey research, methodological practices and substantive findings from U.S. public opinion research are not necessarily transferable to other population contexts. As argued above, this likely also applies to practices and findings when using LLMs, with potentially dire consequences for research results in terms of accuracy and validity, and for social and political reactions to such results in the form of policy or behavior.

[5] In the following chapters, "context" will mainly refer to population contexts, relating to nationality, culture, language, social structure, and political systems. However, "context" can also refer to other aspects, such as task or topic.

Regarding the supplementation or substitution of survey data with LLM-generated "synthetic samples", existing research is contested regarding the U.S. context (e.g., Argyle et al., 2023 vs. Bisbee et al., 2024). Moreover, systematic research extending the scope to other cultural and linguistic contexts has been missing – at the time of writing, there are virtually no studies systematically examining the *joint* impact of language and culture on *individual-level* estimates of public opinion based on LLMs. Findings are thus hardly comparable with U.S.-based studies, and it remains unclear whether the (debatable) learnings from the latter are transferable across contexts, or whether there indeed are representational issues for those contexts. If the attitudes and behaviors of certain population segments are underrepresented in LLM training data and alignment processes, estimates based on such data could be biased. This is likely the case for non-English-speaking contexts featuring a political system more complex than the issue-aligned U.S. two-party system. In Chapters 2 and 3, I investigate the consequences of such potential coverage biases for this application of LLMs in survey research, using the example of vote choice, which is a prominent and challenging aspect of public opinion polling that has experienced drastic mode changes in recent years (Kennedy, Popky, & Keeter, 2023): Based on *native-language, individual-level* profiles with which LLMs are prompted, Chapter 2 features a case study, simulating German vote choice using a single LLM, with individual-level reported voting behavior as the reference. I show how an LLM of one of the industry leaders (OpenAI's GPT) compares to survey-based estimations of public opinion in Germany, and showcase which individual-level factors influence its estimates. I also discuss validity and reliability issues with LLM-synthetic samples. The findings offer important insights for international survey researchers and pollsters, warranting them to be wary of using LLM-synthetic samples. Furthermore, as the example of voting behavior has also been used for testing other new data sources for survey research (e.g., Bach et al., 2021; Behnert et al., 2023; Smith & Gustafson, 2017), the present study could enable researchers to compare the performance of LLM-synthetic sampling to other data sources in the wider polling and election prediction discourse. In addition, Chapter 2 highlights the need for more cross-cultural methodological research in this area, since the results suggest a limited generalizability of U.S.-based studies to the German context and the argumentation presented in the chapter can also be applied beyond Germany – to contexts presenting even more challenging prediction tasks for LLMs by way of linguistic and political-attitudinal data. In Chapter 3, I present such research, extending the case study discussed in Chapter 2 by presenting a comparative design featuring multilingual, pan-European predictions of European elections across a *series of LLMs* (including the successor of the LLM used in Chapter 2), with aggregate national results *unobserved at the time of prediction* as the reference. Beyond showcasing *cross-national and cross-lingual* differences in a truly *pre*dictive task, I also provide insights into the performative differences between proprietary (GPT) and open-source (Llama, Mistral) LLMs. There have been calls in the scientific community for the use of open-source LLMs for research purposes (Barrie et al., 2024; Palmer et al., 2023; Spirling, 2023) due to their advantages regarding transparency, reliability, reproducibility, privacy, and cost. However, there may be trade-offs in terms of performance, as well as expertise and computational resources required (I elaborate on this in Chapter 4). I illustrate how past training and survey data informs LLMs' predictions of future outcomes. By testing several prompt versions, I give insights into how much past information is necessary for accurate prediction, and, by extension, whether high-quality survey data can be re-purposed beyond the initial context of data collection. The study's findings have methodological implications, calling for improving LLMs through, e.g., fine-tuning if they are to be used for such survey research tasks, and for the need of improvement especially for (European) open-source LLMs. In addition, they contribute relevant evidence for the global

polling industry regarding the (lack of) feasibility of predicting elections with synthetic samples based on off-the-shelf LLMs.

Chapter 4 examines an application of LLMs related to the measurement aspect of survey research: the processing of survey data, specifically, coding open-ended responses with LLMs. Research on this topic, while emerging, is sparse. At the time of writing, I am aware of only two published works explicitly using modern-day general-purpose LLMs for coding open-ended survey questions, and they focus on simple classification tasks – English-language data with few categories (Rytting et al., 2023) or broadly discussed topics (Mellon et al., 2024). In this use case, LLMs' general-purpose design, optimized for more common tasks, languages, and topics, could have negative consequences on processing data that does not, for example, contain English-language and U.S.-based political attitudes. Findings of studies using LLMs on other types of social science text data, such as political manifestos or social media data (e.g., Ornstein et al., 2024; Törnberg, 2024), might not be transferable due to the idiosyncratic characteristics of open-ended survey responses – usually being short and topic-specific, but without context. Once again, it is thus unclear how existing findings generalize. I address this gap in Chapter 4, once again relying on German individual-level survey data like in Chapter 2, and using newer versions of the same three LLM families that are tested in Chapter 3. The chapter thereby also gives insights into the (pace of) development of LLMs' capabilities. As suggested in the preceding chapters, I also introduce *fine-tuning* as a potential solution for improving LLMs for survey research applications, that is, further training the LLM on example input for the task at hand and corresponding desired output – here, German responses on survey motivation and their assignment to one of over 20 categories. I compare this approach to the less resource-intensive solution of prompt-tuning, i.e., *few-shot prompting*, where a handful of examples of input and desired output are included in the task description, and to *zero-shot prompting*, where no examples are given. The chapter offers important practical insights for survey researchers, especially those dealing with open-ended response data for a specific, less common topic or language: In order to successfully leverage LLMs for this task, the selection of LLM and prompting approach matters, and researchers need sufficient expertise with fine-tuning LLMs and the computational resources to do so.

Table 1.1 gives an overview of the dimensions studied in this dissertation and their (dis)similarities across chapters.

Consequently, this dissertation makes both methodological and applied contributions to survey research. For survey methodologists, it presents theoretical discussions of types and potential sources of bias in LLM-based survey research from multiple comparative angles and empirically tests their prevalence. For polling practitioners, it showcases concrete applications of LLMs across several steps in the research process and several substantive topics, explaining their practical implementation and highlighting their possibilities and pitfalls. Overall, this dissertation thus provides guidance on challenges regarding the key purpose of survey methodology – ensuring data quality – when using LLMs. Since surveys are a key tool for social science research more broadly, and findings about LLMs in the context of surveys might be transferable to other social science applications, the work presented in this dissertation also contributes to the larger discourse about LLMs as a social science research tool. Conversely, it also contributes to the understanding and mitigation of biases in LLMs, thereby aiding computational (social) scientists in improving them. The next section gives a more detailed summary of the following chapters.

Dimension	Chapter 2	Chapter 3	Chapter 4
Application (Substantive Topic)	Simulating respondents (Predicting vote choice)		Coding open-ended responses (Survey motivation)
Scope (Countries / Languages)	Case study (DE)	Comparative (EU)	Case study (DE)
Perspective	Retrospective	Prospective	Retrospective
LLMs	GPT-3.5	GPT-4 Turbo, Llama-3.1, Mistral-7b	GPT-4o, Llama-3.2, Mistral Nemo
Prompting Approaches	zero-shot	zero-shot (2 prompt versions)	zero-shot (2 prompt versions), few-shot, fine-tuned
Level of Analysis	Individual	Aggregate	

Table 1.1: Analytical dimensions covered in this dissertation.

1.5 Summary of Chapters

Chapter 2: Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice[6]

In Chapter 2, I address one of the most prominent discussions in the nascent subfield of LLM-based survey research: Using LLM-generated "synthetic samples" to complement or replace traditional surveys, being much more cost- and time-efficient than the latter. Proponents of this method base it on the assumption that LLMs were trained on human-generated text, therefore potentially reflecting human attitudes and behavior. While the application studied in Chapter 2 (and 3) can be used in several stages of the survey research process (as detailed earlier), the challenge in terms of survey error always lies with coverage.

As mentioned earlier, a number of mostly U.S.-based studies have prompted LLMs to mimic survey respondents, with initial studies finding that the responses closely match the survey data (Argyle et al., 2023). Other studies contest these findings in the U.S. context (e.g., Bisbee et al., 2024). However, regardless of success, several contextual factors related to the relationship between the respective target population and LLM training data and alignment processes might affect the generalizability of such findings. These factors are related to the digital divide, i.e., the (lack of) coverage of the target population in the training data and alignment processes in terms of the language, attitudes, and relationships between those attitudes and individual- and country-level characteristics. In this chapter, I investigate the cross-cultural and cross-lingual generalizability of early U.S.-based findings. I test to what extent LLMs can estimate vote choice in Germany on aggregate and for different population subgroups. I choose this outcome because of its relevance in the societal and scientific discourse, the challenge it poses for survey researchers and pollsters (Kennedy, Blumenthal, et al., 2023), and its strong dependency on national social, political, linguistic, and attitudinal context related to the arguments against the generalizability of LLM-synthetic sampling applications (e.g., Dalton, 2018; Ford & Jennings, 2020; Inglehart,

[6] A journal article version of this chapter has been published as von der Heyde, L., Haensch, A.-C., & Wenz, A. (2025). Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice. *Social Science Computer Review*, 0(0). https://doi.org/10.1177/08944393251337014.

1977; G. Jansen et al., 2013; Lipset & Rokkan, 1967; Sass & Kuhnle, 2023). At the same time, choosing vote choice and Germany as test cases presents a middle ground for examining LLM-based synthetic samples, being informative for (even) lower-resource topics and populations in LLM training and alignment.

To generate a synthetic sample of eligible voters in Germany, I create "personas", i.e., small profiles, matching the individual characteristics of the 2017 German Longitudinal Election Study respondents (GLES, 2019). I chose this dataset for its high-quality probability design, and the fact that the 2017 election occurred before the training data cutoff for the selected LLM (GPT-3.5) – information about the election is likely included in that training data, making it an easy test case. Any limitations found for such a "medium-resource" context will likely extend to, if not be intensified, in temporal contexts outside of the training data window. The personas include information known to be associated with differences in voting behavior – demographics, party affiliations, and attitudes on salient issues. Prompting GPT-3.5 with each persona in German, I ask the LLM to predict each respondents' vote choice in the 2017 German federal elections and compare these predictions to the survey-based estimates on the aggregate and subgroup levels. Thus, the purpose of this study (in contrast to the one presented in Chapter 3) is not to compare LLM-based estimates of vote choice to actual election results, but to assess whether LLMs can infer individual voting behavior and arrive at estimates comparable to those made with individual-level survey data. 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.

I find that GPT-3.5 does not predict citizens' vote choice accurately, exhibiting a bias towards the Green and Left parties, and making better predictions for more "typical" voter subgroups. While the language model is able to capture broad-brush tendencies tied to partisanship, it tends to miss out on the multifaceted factors that sway individual voter choices. As a consequence, not only are LLM-synthetic samples not helpful for estimating how groups likely swinging an election, such as non-partisans, will vote, they also risk underestimating the popularity of parties without a strong partisan base. Such samples thus provide little added value over survey-based estimates. Furthermore, the results suggest that GPT-3.5 might not be reliable for estimating nuanced, subgroup-specific political attitudes. I also discuss the implications of these findings regarding the disparities in opinion representation in LLMs and the limitation of applying them for public opinion estimation more broadly when not accounting for the biases in their training data and alignment processes.

Chapter 3: United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections[7]

In Chapter 3, I apply the argument on context-dependent biases in LLM-based survey research outlined in Chapter 2 beyond Germany. As described before, Germany might present a medium-resource test case, with other contexts having even lower prevalence in LLM training and alignment processes. Cross-national research on attitudinal biases in LLMs has typically employed

[7] A journal article version of this chapter is currently under review: von der Heyde, L., Haensch, A.-C., Wenz, A., & Ma, B. *United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections.* A previous version has been published as a preprint on arXiv at https://doi.org/10.48550/arXiv.2409.09045.

country-level information only (e.g., Durmus et al., 2024). Research on estimating public opinion with LLM-synthetic samples based on individual-level characteristics that has identified biases regarding certain subgroups (e.g., Bisbee et al., 2024), in contrast, has mostly been conducted in isolated national settings thus far. Additionally, it would be advantageous if LLM-synthetic samples could be used for making accurate predictions of future outcomes. Yet, the focus of existing research largely has been on "predicting the past". U.S.-based experiments on prediction have either yielded worse results than retrodiction (J. Kim & Lee, 2023) or, in the case of predicting the 2024 U.S. presidential elections, failed (Mendoza, 2024, but see Jiang et al., 2024). This should not be surprising considering the temporal constraints of LLMs and the inherently challenging task of predicting elections even with other methods (this being both the reason for researchers and startups turning to LLMs for solving this challenge, and for the observed failure to do so), but warrants rigorous research. I therefore extend the test of LLM-based, individual-level, native-language predictions of public opinion to the entire European Union. More specifically, I examine to what extent LLM-based predictions of individual voting behavior exhibit context-dependent biases by predicting the results of the 2024 European Parliament elections.

To do so, I once again create personas containing socio-demographic and attitudinal information based on a probability-based survey, the Eurobarometer (EB 99.4, European Commission, 2024), including 26,000 eligible voters in all 27 European Union (EU) member states. Importantly, as the survey data was collected a year before the elections took place (but before the LLM training data cutoff), it does not contain information on voting behavior or intention. Instead, the predictions can only be compared to the actual election results in aggregated form. To understand whether any biases found generalize across LLMs, I compare the proprietary LLM GPT-4-Turbo with the open-source LLMs Llama-3.1 and Mistral, always ensuring privacy by only providing anonymized profiles and by hosting the LLMs on secure, European servers. A week before the European elections in June 2024, I prompted the LLMs with the personas and asked them to predict each person's voting behavior. For all countries, I prompted the LLMs with the individual profiles in English, once containing only socio-demographic information, and once also containing the attitudinal variables. For an additional in-depth investigation of differences in LLMs' bias across languages, I selected six linguistically, socio-structurally, and politically diverse countries – France, Germany, Ireland, Poland, Slovakia, and Sweden – for which I prompted the LLMs in the respective country's native language.

After the elections' conclusion, I compare the aggregate predicted party vote shares to the official national-level results for each country, differentiating between turnout and party vote shares among voters. Beyond contrasting the differences across all countries, I analyze the LLMs' predictive performance based on prompts in English and the six selected countries' native languages, as well as differences in predictive performance depending on the amount and kind of individual-level information contained in the prompt. I show that LLM-based predictions of future voting behavior largely fail – they overestimate turnout and are largely unable to accurately predict the winner, rank ordering, or individual party vote shares. Only providing socio-demographic information about individual voters further worsens the results, casting doubts on the feasibility of using LLM-based synthetic samples as a supplement or substitution of detailed survey data. Finally, LLMs' predictive accuracy is unequally distributed across national and linguistic contexts. LLMs are especially bad at predicting voting behavior for Eastern European countries and countries with Slavic native languages, regardless of language used or the amount of information provided in the prompt, suggesting systematic contextual biases. These findings emphasize the limited applicability of LLM-synthetic samples to public opinion prediction across contexts. I discuss the differences between LLMs and public opinion polls with regards to the purpose they

were designed for and their temporal constraints. Without further adaptation through, e.g., fine-tuning with more recent and target-population-specific public opinion data, off-the-shelf LLMs appear infeasible for public opinion prediction not just in terms of accuracy, but also in terms of efficiency, highlighting a trade-off between the recency and level of detail of available survey data for synthetic samples.

Chapter 4: AIn't Nothing But a Survey? Using Large Language Models for Coding German Open-Ended Survey Responses on Survey Motivation[8]

In Chapter 4, I turn to the use of LLMs in the data processing stage of the survey life cycle. Specifically, I investigate their usability for classifying open-ended survey responses and the potential processing errors LLMs could induce in terms of accuracy and reliability. Due to their linguistic capacities, it is likely that LLMs are an efficient alternative to time-consuming manual coding and the pre-training of supervised machine learning models. As the sparse existing studies on this topic have focused on English-language responses relating to non-complex topics or on single LLMs, it is unclear whether their findings generalize and how the quality of such classifications compares to established methods. Moreover, research on LLM-based classification of social science texts more broadly shows competing evidence regarding cross-lingual performance, differences between LLMs, and prompting approaches. Finally, as highlighted earlier, open-ended survey responses constitute a very specific type of natural language text, often being quite short and lacking context (see, e.g., Schonlau et al., 2023). Thus, findings related to other types of social science text might not be transferable to open-ended survey responses.

In this study, I test to what extent different LLMs can be used to code German open-ended survey responses on a specific and complex topic. Once again, I compare the most recently available LLMs of the GPT, Llama, and Mistral families, thereby providing insights into the performance differences of open- and closed source LLMs. Furthermore, I compare several prompting approaches, including zero- and few-shot prompting and fine-tuning. For this investigation, I use a sample of 5072 open-ended responses on survey motivation (i.e., reasons why respondents participate in the survey) from the GESIS Panel.pop Population Sample (Bosnjak et al., 2018; GESIS, 2024). Beyond the advantageous linguistic and topical specificity, the dataset contains uni-dimensional responses and verified human expert codes, making it a convenient test case. I prompt the LLMs in German with a predefined coding scheme and instruct them to classify each survey response, either based on the coding scheme alone, when adding descriptions or examples for each category, or when fine-tuning the LLM with a subset of response-category pairs. I evaluate the LLMs' performance by comparing its classifications to those made by the human coders.

While the tested LLMs appear reliable in the short-term, only fine-tuning achieves satisfactory levels of predictive accuracy that are comparable to supervised methods. Performance differences between prompting approaches are conditional on the LLM used, as overall performance differs greatly between LLMs: GPT performs best in terms of accuracy, and few-shot prompting leads to the best performance. Disregarding fine-tuning, the prompting approach is not as important when using GPT, but makes a big difference for other LLMs, especially Mistral.

[8] A journal article version of this chapter has been accepted at *Survey Research Methods*: von der Heyde, L., Haensch, A-.C., Weiß, B., & Daikeler, J. (forthcoming). *Using Large Language Models for Coding German Open-Ended Survey Responses on Survey Motivation.* A previous version has been published as a preprint on arXiv at https://doi.org/10.48550/arXiv.2506.14634.

Finally, LLMs' unequal classification performance across different categories results in different categorical distributions when not using fine-tuning. In particular, the LLMs struggle with non-substantive catch-all categories, which tend to be common in open-ended responses. In sum, the applicability of LLMs for coding open-ended responses not only depends on the LLM and prompting approach used, but also on the topic (in terms of specificity and categorical complexity) and possibly the language of the responses.

I discuss the implications of these findings, both for methodological research on coding open-ended responses (the need for LLMs to be fine-tuned for this task making them not the resource-efficient, easily accessible alternative researchers may have hoped) and for their substantive analysis (the use of LLM-generated codes potentially leading to a different understanding of the concept being measured), and the many trade-offs researchers need to consider when choosing automated methods for open-ended response classification in the age of LLMs.

Chapter 5: Discussion and Conclusion

I conclude this dissertation by summarizing its main findings and discussing them in the light of the ongoing developments in the rapidly evolving LLM research landscape. I also point to avenues for future work.

References

Adhikari, D. M., Cannanure, V. K., Hartland, A., & Weber, I. (2025). Exploring LLMs for Automated Pre-Testing of Cross-Cultural Surveys. http://arxiv.org/abs/2501.05985v1

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The Illusion of Artificial Inclusion [event-place: Honolulu, HI, USA]. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642703

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies [Place: Honolulu, Hawaii, USA]. *Proceedings of the 40th International Conference on Machine Learning*.

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2025). Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 15–36. https://doi.org/10.1609/icwsm.v19i1.35801

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Anthropic. (2025). *Claude 3.7 Sonnet System Card* (tech. rep.). Retrieved March 26, 2025, from https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 1–15. https://doi.org/10.1017/pan.2023.2

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023, September). Which Humans? https://doi.org/10.31234/osf.io/5b26t

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting Voting Behavior Using Digital Trace Data. *Social Science Computer Review*, *39*(5), 862–883. https://doi.org/10.1177/0894439319882896

Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Barari, S., Slowinski, Z., Wang, N., Angbazo, J., Sepulvado, B., Christian, L., & Dean, E. (2024, November). *Generative AI Can Enhance Survey Interviews* (tech. rep.). NORC at the University of Chicago. Retrieved February 26, 2025, from https://www.norc.org/research/library/generative-ai-can-enhance-survey-interviews.html

Barrie, C., Palmer, A., & Spirling, A. (2024). Replication for Language Models: Problems, Principles, and Best Practice for Political Science. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Batzner, J., Stocker, V., Schmid, S., & Kasneci, G. (2024, July). GermanPartiesQA: Benchmarking Commercial Large Language Models for Political Bias and Sycophancy [arXiv:2407.18008 [cs]]. Retrieved September 18, 2024, from http://arxiv.org/abs/2407.18008

Behnert, J., Lajic, D., & Bauer, P. C. (2023, November). *Can we predict multi-party elections with Google Trends data? Evidence across elections, data windows, and model classes* (preprint). Open Science Framework. https://doi.org/10.31219/osf.io/6duw2

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

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. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 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*, 1–16. https://doi.org/10.1017/pan.2024.5

Blank, G. (2013). WHO CREATES CONTENT?: Stratification and content creation on the Internet. *Information, Communication & Society*, *16*(4), 590–612. https://doi.org/10.1080/1369118X.2013.777758

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel. *Social Science Computer Review*, *36*(1), 103–115. https://doi.org/10.1177/0894439317697949

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

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

Buschek, C., & Thorp, J. (n.d.). Models All The Way Down. Retrieved February 26, 2025, from https://knowingmachines.org/models-all-the-way

Carrière, T. C., Boeschoten, L., Struminskaya, B., Janssen, H. L., De Schipper, N. C., & Araujo, T. (2024). Best practices for studies using digital data donation. *Quality & Quantity*. https://doi.org/10.1007/s11135-024-01983-x

Cerina, R., & Duch, R. (2023, September). Artificially Intelligent Opinion Polling [arXiv:2309.06029 [stat]]. Retrieved September 21, 2023, from http://arxiv.org/abs/2309.06029

Clemmensen, L. H., & Kjærsgaard, R. D. (2023, February). Data Representativity for Machine Learning and AI Systems [arXiv:2203.04706 [cs, stat]]. Retrieved April 4, 2023, from http://arxiv.org/abs/2203.04706

Conrad, F. G., Keusch, F., & Schober, M. F. (2021). New Data in Social and Behavioral Research. *Public Opinion Quarterly*, *85*(S1), 253–263. https://doi.org/10.1093/poq/nfab027

Conrad, F. G., Schober, M., Nielsen, D., & Reichert, H. (2019). Race-of-Virtual-Interviewer Effects. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1011&context=sociw

Conrad, F. G., Schober, M. F., Jans, M., Orlowski, R. A., Nielsen, D., & Levenstein, R. (2015). Comprehension and engagement in survey interviews with virtual agents. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01578

Couper, M. P. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, *7*(3), 145–156. https://doi.org/https://doi.org/10.18148/srm/2013.v7i3.5751

Crockett, M., & Messeri, L. (2023, June). Should large language models replace human participants? https://doi.org/10.31234/osf.io/4zdx9

Cuevas, A., Brown, E. M., Scurrell, J. V., Entenmann, J., & Daepp, M. I. G. (2023, October). Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models [arXiv:2309.10187 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2309.10187

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. (2018, September). *Political Realignment: Economics, Culture, and Electoral Change* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780198830986.001.0001

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025, January). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [arXiv:2501.12948 [cs]]. https://doi.org/10.48550/arXiv.2501.12948

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00241-5

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, *120*(51), e2309583120. https://doi.org/10.1073/pnas.2309583120

Dey, K., Tarannum, P., Hasan, M. A., Razzak, I., & Naseem, U. (2024, October). Better to Ask in English: Evaluation of Large Language Models on English, Low-resource and Cross-Lingual Settings [arXiv:2410.13153 [cs] version: 1]. https://doi.org/10.48550/arXiv.2410.13153

D'Ignazio, C., & Klein, L. F. (2020, March). *Data Feminism* [_eprint: https://direct.mit.edu/book-pdf/2390355/book_9780262358521.pdf]. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., … Zhao, Z. (2024, August). The Llama 3 Herd of Models [arXiv:2407.21783 [cs]]. https://doi.org/10.48550/arXiv.2407.21783

Durmus, E., Nguyen, K., Liao, T., 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. (2024). Towards Measuring the Representation of Subjective Global Opinions in Language Models. *First Conference on Language Modeling*. https://openreview.net/forum?id=zl16jLb91v

Dutwin, D., & Buskirk, T. D. (2023). A Deeper Dive into the Digital Divide: Reducing Coverage Bias in Internet Surveys. *Social Science Computer Review*, *41*(5), 1902–1920. https://doi.org/10.1177/08944393221093467

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs [_eprint: https://www.science.org/doi/pdf/10.1126/science.adj0998]. *Science*, *384*(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

European Commission, B. (2024). Eurobarometer 99.4 (2023) [Published: GESIS, Cologne. ZA7997 Data file Version 1.0.0, https://doi.org/10.4232/1.14167]. https://doi.org/10.4232/1.14167

Foerderer, J. (2023, August). Should we trust web-scraped data? [arXiv:2308.02231 [cs, econ, q-fin, stat]]. Retrieved August 18, 2023, from http://arxiv.org/abs/2308.02231

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

Fröhling, L., Demartini, G., & Assenmacher, D. (2024, October). Personas with Attitudes: Controlling LLMs for Diverse Data Annotation [arXiv:2410.11745]. Retrieved October 23, 2024, from http://arxiv.org/abs/2410.11745

Geisen, E. (2024). Prompting Insight: Enhancing Open-Ended Survey Responses with AI-Powered Follow-Ups. *79th Annual AAPOR Conference*. https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/3103

GESIS. (2024). GESIS Panel - Extended Edition [Published: GESIS, Cologne. ZA5664 Data file Version 54.0.0, https://doi.org/10.4232/1.14385]. https://doi.org/10.4232/1.14385

Ghosh, S., & Caliskan, A. (2023). ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages [arXiv:2305.10510 [cs]]. *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, 397–415. https://doi.org/10.1145/3568813.3600120

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks [arXiv:2303.15056 [cs]]. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

GLES. (2019). Post-election Cross Section (GLES 2017) Nachwahl-Querschnitt (GLES 2017). 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*. https://doi.org/10.1037/met0000540

Gross, N. (2023). What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI. *Social Sciences*, *12*(8), 435. https://doi.org/10.3390/socsci12080435

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108–1109. https://doi.org/10.1126/science.adi1778

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. John Wiley & Sons.

Gui, G., & Toubia, O. (2023). The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4650172

Gummer, T., Höhne, J. K., Rettig, T., Roßmann, J., & Kummerow, M. (2023). Is there a growing use of mobile devices in web surveys? Evidence from 128 web surveys in Germany. *Quality & Quantity*, *57*(6), 5333–5353. https://doi.org/10.1007/s11135-022-01601-8

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2024, January). Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs [arXiv:2311.04892 [cs]]. https://doi.org/10.48550/arXiv.2311.04892

Hada, R., Seth, A., Diddee, H., & Bali, K. (2023, December). "Fifty Shades of Bias": Normative Ratings of Gender Bias in GPT Generated English Text. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1862–1876). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.115

Haim, A., Salinas, A., & Nyarko, J. (2024, February). What's in a Name? Auditing Large Language Models for Race and Gender Bias [arXiv:2402.14875 [cs]]. https://doi.org/10.48550/arXiv.2402.14875

Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 63–76. https://doi.org/10.1177/0002716215570866

Hargittai, E. (2020). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Harkness, J. (2003). Questionnaire translation. In *Cross-Cultural Survey Methods – J. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.)* (pp. 35–56). Wiley.

Hartmann, J., Schwenzow, J., & Witte, M. (2023, January). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation [arXiv:2301.01768 [cs]]. Retrieved March 28, 2023, from http://arxiv.org/abs/2301.01768

Havaldar, S., Singhal, B., Rai, S., Liu, L., Guntuku, S. C., & Ungar, L. (2023, July). Multilingual Language Models are not Multicultural: A Case Study in Emotion. In J. Barnes, O. De Clercq, & R. Klinger (Eds.), *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 202–214). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wassa-1.19

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

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text [Publisher: SAGE Publications Ltd]. *Research & Politics*, *11*(1), 20531680241236239. https://doi.org/10.1177/20531680241236239

Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R. (2024). Predicting Results of Social Science Experiments Using Large Language Models.

Hoffmann, C. P., Lutz, C., & Meckel, M. (2015). Content creation on the Internet: A social cognitive perspective on the participation divide. *Information, Communication & Society*, *18*(6), 696–716. https://doi.org/10.1080/1369118X.2014.991343

Höhne, J. K., Claassen, J., & Wolf, B. L. (2025). LLM-driven bot infiltration: Protecting web surveys through prompt injections. Retrieved March 20, 2025, from https://jkhoehne.eu/wp-content/uploads/2025/02/hoehne-et-al-2025-LLM-driven-bot-infiltration-preprint-1.pdf

Hölzl, J., Keusch, F., & Sajons, C. (2025). The (mis)use of Google Trends data in the social sciences - A systematic review, critique, and recommendations. *Social Science Research*, *126*, 103099. https://doi.org/10.1016/j.ssresearch.2024.103099

Hommel, B. E. (2023). Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. *Personality and Individual Differences*, *213*, 112307. https://doi.org/10.1016/j.paid.2023.112307

Horton, J. J. (2023, April). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* (Working Paper No. 31122) (Series: Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w31122

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12432]. *Language and Linguistics Compass*, *15*(8), e12432. https://doi.org/10.1111/lnc3.12432

Huckle, J., & Williams, S. (2025). Easy Problems that LLMs Get Wrong. In K. Arai (Ed.), *Advances in Information and Communication* (pp. 313–332). Springer Nature Switzerland.

Iglesias, P. A., Ochoa, C., & Revilla, M. (2024). A practical guide to (successfully) collect and process images through online surveys. *Social Sciences & Humanities Open*, *9*, 100792. https://doi.org/10.1016/j.ssaho.2023.100792

Inglehart, R. (1977). *The Silent Revolution: Changing Values and Political Styles Among Western Publics* [OCLC: 979580560]. Princeton University Press.

International Telecommunication Union. (2022). *Measuring digital development Facts and Figures 2022* (tech. rep.).

Jacobsen, R. M., Cox, S. R., Griggio, C. F., & Van Berkel, N. (2025). Chatbots for Data Collection in Surveys: A Comparison of Four Theory-Based Interview Probes. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3706598.3714128

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, B. J., Jung, S.-g., & Salminen, J. (2023). Employing large language models in survey research. *Natural Language Processing Journal*, *4*, 100020. https://doi.org/10.1016/j.nlp.2023.100020

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 [Number: 2]. *Social Science Research*, *42*(2), 376–400. https://doi.org/10.1016/j.ssresearch.2012.09.007

Jiang, S., Wei, L., & Zhang, C. (2024, November). Donald Trumps in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models [arXiv:2411.01582 [econ]]. Retrieved November 19, 2024, from http://arxiv.org/abs/2411.01582

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022, March). The Ghost in the Machine has an American accent: Value conflict in GPT-3 [arXiv:2203.07785 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2203.07785

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2017). Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support. *Social Science Computer Review*, *35*(3), 336–356. https://doi.org/10.1177/0894439316631043

Kalinin, K. (2023). Improving GPT Generated Synthetic Samples with Sampling-Permutation Algorithm. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4548937

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., Mcgeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., Witt, E., & Wlezien, C. (2023). *AN EVALUATION OF 2016 ELECTION POLLS IN THE UNITED STATES* (tech. rep.). American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2023/01/AAPOR-2016-Election-Polling-Report.pdf

Kennedy, C., Popky, D., & Keeter, S. (2023, April). *How Public Polling Has Changed in the 21st Century* (tech. rep.). Pew Research Center. https://www.pewresearch.org/methods/2023/04/19/how-public-polling-has-changed-in-the-21st-century/

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

Kim, J., & Lee, B. (2023, November). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction [arXiv:2305.09620 [cs]]. Retrieved January 23, 2024, from http://arxiv.org/abs/2305.09620

Kohler, U. (2020). Survey Research Methods during the COVID-19 Crisis [Artwork Size: 93-94 Pages Publisher: Survey Research Methods]. *Survey Research Methods*, 93–94 Pages. https://doi.org/10.18148/SRM/2020.V14I2.7769

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*. https://doi.org/10.1007/s43681-023-00312-6

Kreuter, F. (2025). Modernizing Data Collection. *Journal of Official Statistics*, *41*(3), 863–872. https://doi.org/10.1177/0282423X251318452

Kuntz, J. B., & Silva, E. C. (2023, September). *Who Authors the Internet? Analyzing Gender Diversity in ChatGPT-3 Training Data* (tech. rep.). Pitt Cyber – Institute for Cyber Law, Policy, and Security.

Lang, M. M., & Eskenazi, S. (2025, February). Telephone Surveys Meet Conversational AI: Evaluating a LLM-Based Telephone Survey System at Scale [arXiv:2502.20140 [cs]]. https://doi.org/10.48550/arXiv.2502.20140

Laverghetta Jr., A., & Licato, J. (2023, July). Generating Better Items for Cognitive Assessments Using Large Language Models. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 414–428). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.34

Lebrun, B., Temtsin, S., Vonasch, A., & Bartneck, C. (2024). Detecting the corruption of online questionnaires by artificial intelligence. *Frontiers in Robotics and AI, Volume 10 - 2023*. https://doi.org/10.3389/frobt.2023.1277635

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

Lerner, J. (2024, October). The Promise & Pitfalls of AI-Augmented Survey Research. Retrieved March 11, 2025, from https://www.norc.org/research/library/promise-pitfalls-ai-augmented-survey-research.html

Li, B., Haider, S., & Callison-Burch, C. (2024). This Land is Your, My Land: Evaluating Geopolitical Bias in Language Models through Territorial Disputes. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3855–3871. https://doi.org/10.18653/v1/2024.naacl-long.213

Liew, A., & Mueller, K. (2022, December). Using Large Language Models to Generate Engaging Captions for Data Visualizations [arXiv:2212.14047 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2212.14047

Linegar, M., Kocielnik, R., & Alvarez, R. M. (2023). Large language models and political science. *Frontiers in Political Science*, *5*, 1257092. https://doi.org/10.3389/fpos.2023.1257092

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.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts [_eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00638/2336043/tacl_a_00638.pdf]. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Luiten, A., Hox, J., & Leeuw, E. d. (2020). Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys [_eprint: https://doi.org/10.2478/jos-2020-0025]. *Journal of Official Statistics*, *36*(3), 469–487. https://doi.org/10.2478/jos-2020-0025

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

Lyberg, L., Stange, M., Harkness, J., Mohler, P., Pennell, B.-E., & Japec, L. (2014, August). A review of quality issues associated with studying hard-to-survey populations. In R. Tourangeau, B. Edwards, T. P. Johnson, K. M. Wolter, & N. Bates (Eds.), *Hard-to-Survey Populations* (1st ed., pp. 82–108). Cambridge University Press. https://doi.org/10.1017/CBO9781139381635.007

Ma, W., Chiang, B., Wu, T., Wang, L., & Vosoughi, S. (2023). Intersectional Stereotypes in Large Language Models: Dataset and Analysis. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8589–8597. https://doi.org/10.18653/v1/2023.findings-emnlp.575

Maiorino, A., Padgett, Z., Wang, C., Yakubovskiy, M., & Jiang, P. (2023). Application and Evaluation of Large Language Models for the Generation of Survey Questions. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 5244–5245. https://doi.org/10.1145/3583780.3615506

Masoud, R., Liu, Z., Ferianc, M., Treleaven, P. C., & Rodrigues, M. R. (2025, January). Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 8474–8503). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.567/

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023, September). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve [arXiv:2309.13638 [cs]]. Retrieved September 18, 2024, from http://arxiv.org/abs/2309.13638

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, *11*(1). https://doi.org/10.1177/20531680241231468

Mendoza, D. (2024). AI polling company defends wrong predictions on the US election. *Semafor*. Retrieved November 21, 2024, from https://www.semafor.com/article/11/06/2024/ai-startup-aaru-defends-using-artificial-intelligence-for-polling

Moore, J., Deshpande, T., & Yang, D. (2024, July). Are Large Language Models Consistent over Value-laden Questions? [arXiv:2407.02996 [cs]]. Retrieved July 30, 2024, from http://arxiv.org/abs/2407.02996

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

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M., & Harwood, P. (2014). Social Media in Public Opinion Research: Executive Summary of the Aapor Task Force on Emerging Technologies in Public Opinion Research. *Public Opinion Quarterly*, *78*(4), 788–794. https://doi.org/10.1093/poq/nfu053

Nagireddy, M., Chiazor, L., Singh, M., & Baldini, I. (2024). SocialStigmaQA: A Benchmark to Uncover Stigma Amplification in Generative Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(19), 21454–21462. https://doi.org/10.1609/aaai.v38i19.30142

Nie, M. (2024). Artificial Intelligence: The Biggest Threat to Democracy Today? *Proceedings of the AAAI Symposium Series*, *3*(1), 376–379. https://doi.org/10.1609/aaaiss.v3i1.31239

Niszczota, P., Janczak, M., & Misiak, M. (2025). Large language models can replicate cross-cultural differences in personality. *Journal of Research in Personality*, *115*, 104584. https://doi.org/10.1016/j.jrp.2025.104584

Olivos, F., & Liu, M. (2024). ChatGPTest: Opportunities and Cautionary Tales of Utilizing AI for Questionnaire Pretesting [_eprint: https://doi.org/10.1177/1525822X241280574]. *Field Methods*, *0*(0), 1525822X241280574. https://doi.org/10.1177/1525822X241280574

OpenAI. (2024). *Sora System Card* (tech. rep.). Retrieved March 26, 2025, from https://openai.com/index/sora-system-card/

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 Number: 6]. https://doi.org/10.48550/ARXIV.2303.08774

Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2024). How to Train Your Stochastic Parrot: Large Language Models for Political Texts. https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf

Ostrow, R., & Lopez, A. (2025, January). LLMs Reproduce Stereotypes of Sexual and Gender Minorities [arXiv:2501.05926 [cs] version: 1]. https://doi.org/10.48550/arXiv.2501.05926

Öztürk, I. T., Nedelchev, R., Heumann, C., Arias, E. G., Roger, M., Bischl, B., & Aßenmacher, M. (2025). How Different is Stereotypical Bias Across Languages? In R. Meo & F. Silvestri (Eds.), *Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (pp. 209–229). Springer Nature Switzerland.

Palacios Barea, M. A., Boeren, D., & Ferreira Goncalves, J. F. (2023). At the intersection of humanity and technology: A technofeminist intersectional critical discourse analysis of gender and race biases in the natural language processing model GPT-3. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01804-z

Palmer, A., Smith, N. A., & Spirling, A. (2023). Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, *4*(1), 2–3. https://doi.org/10.1038/s43588-023-00585-1

Pezeshkpour, P., & Hruschka, E. (2024, June). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006–2017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.130

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2024, June). TopicGPT: A Prompt-based Topic Modeling Framework. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2956–2984). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.164

Qu, Y., & Wang, J. (2024). Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, *11*(1), 1095. https://doi.org/10.1057/s41599-024-03609-x

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. https://cdn.openai.com/papers/whisper.pdf

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021, July). Zero-Shot Text-to-Image Generation. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 8821–8831, Vol. 139). PMLR. https://proceedings.mlr.press/v139/ramesh21a.html

Ramezani, A., & Xu, Y. (2023, July). Knowledge of cultural moral norms in large language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 428–446). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.26

Rettenberger, L., Reischl, M., & Schutera, M. (2025). Assessing political bias in large language models. *Journal of Computational Social Science*, *8*(2), 42. https://doi.org/10.1007/s42001-025-00376-w

Revilla, M., Ochoa, C., Höhne, J. K., & Couper, M. P. (2025). Transcribing and Coding Voice Answers Obtained in Web Surveys: Comparing Three Leading Automatic Speech Recognition Tools and Human versus LLM-based Coding [Publisher: Unpublished]. https://doi.org/10.13140/RG.2.2.15968.39681

Roberts, G. (2022, December). AI Training Datasets: The Books1+Books2 that Big AI eats for breakfast [Section: AI Reality 2022]. Retrieved February 26, 2025, from https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/

Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T. M., & Stern, M. J. (2015). Digital inequalities and why they matter. *Information, Communication & Society*, *18*(5), 569–582. https://doi.org/10.1080/1369118X.2015.1012532

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023, June). Towards Coding Social Science Datasets with Language Models [arXiv:2306.02177 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2306.02177

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sanders, N. E., Ulinich, A., & Schneier, B. (2023). Demonstrations of the Potential of AI-based Political Issue Polling [Publisher: The MIT Press]. *Harvard Data Science Review*, *5*(4).

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023, July). Whose Opinions Do Language Models Reflect? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 29971–30004, Vol. 202). PMLR. https://proceedings.mlr.press/v202/santurkar23a.html

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21982]. *Psychology & Marketing*, *41*(6), 1254–1270. https://doi.org/10.1002/mar.21982

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

Schonlau, M., Weiß, J., & Marquardt, J. (2023). Multi-label classification of open-ended questions with BERT. *2023 Big Data Meets Survey Science (BigSurv)*, 1–8. https://doi.org/10.1109/BigSurv59479.2023.10486634

Schott, T., Furman, D., & Bhat, S. (2023). Polyglot or Not? Measuring Multilingual Encyclopedic Knowledge in Foundation Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11238–11253. https://doi.org/10.18653/v1/2023.emnlp-main.691

Schumacher, S., & Kent, N. (2020, April). 8 charts on internet use around the world as countries grapple with COVID-19. Retrieved October 17, 2023, from https://www.pewresearch.org/short-reads/2020/04/02/8-charts-on-internet-use-around-the-world-as-countries-grapple-with-covid-19/

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A TOTAL ERROR FRAMEWORK FOR DIGITAL TRACES OF HUMAN BEHAVIOR ON ONLINE PLATFORMS. *Public Opinion Quarterly*, *85*, 399–422. https://doi.org/10.1093/poq/nfab018

Shaw, A., & Hargittai, E. (2018). The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing. *Journal of Communication*, *68*(1), 143–168. https://doi.org/10.1093/joc/jqx003

Simmons, G., & Hare, C. (2023). Large Language Models as Subpopulation Representative Models: A Review [https://arxiv.org/abs/2310.17888]. https://doi.org/10.48550/arXiv.2310.17888

Smith, B. K., & Gustafson, A. (2017). Using Wikipedia to Predict Election Outcomes. *Public Opinion Quarterly*, *81*(3), 714–735. https://doi.org/10.1093/poq/nfx007

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science [Bandiera_abtest: a Cg_type: World View Publisher: Nature Publishing Group Subject_term: Ethics, Machine learning, Technology, Scientific community]. *Nature*, *616*(7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4

Struminskaya, B., Lugtig, P., Keusch, F., & Höhne, J. K. (2020). Augmenting Surveys With Data From Sensors and Apps: Opportunities and Challenges. *Social Science Computer Review*, 0894439320979951. https://doi.org/10.1177/0894439320979951

Sultanum, N., & Srinivasan, A. (2023). DATATALES: Investigating the use of Large Language Models for Authoring Data-Driven Articles. *2023 IEEE Visualization and Visual Analytics (VIS)*, 231–235. https://doi.org/10.1109/VIS54172.2023.00055

Tavus. (2025). The OS for Human-AI Interaction. Retrieved April 8, 2025, from https://www.tavus.io/

Tewari, T., & Hosein, P. (2024). Automating the Conducting of Surveys Using Large Language Models. In A. Fred, A. Hadjali, O. Gusikhin, & C. Sansone (Eds.), *Deep Learning Theory and Applications* (pp. 136–151). Springer Nature Switzerland.

Thirunavukarasu, A. J., & O'Logbon, J. (2024). The potential and perils of generative artificial intelligence in psychiatry and psychology. *Nature Mental Health*, *2*(7), 745–746. https://doi.org/10.1038/s44220-024-00257-7

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., & Neubig, G. (2024). Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design [_eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00685/2468689/tacl_a_00685.pdf]. *Transactions of the Association for Computational Linguistics*, *12*, 1011–1026. https://doi.org/10.1162/tacl_a_00685

Törnberg, P. (2024). Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages [Publisher: SAGE Publications Inc]. *Social Science Computer Review*, 08944393241286471. https://doi.org/10.1177/08944393241286471

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

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

Velez, Y. R. (2025). Crowdsourced Adaptive Surveys. *Political Analysis*, 1–14. https://doi.org/10.1017/pan.2024.34

Veselovsky, V., Horta Ribeiro, M., Cozzolino, P. J., Gordon, A., Rothschild, D., & West, R. (2025). Prevalence and Prevention of Large Language Model Use in Crowd Work [Place: New York, NY, USA Publisher: Association for Computing Machinery]. *Commun. ACM*, *68*(3), 42–47. https://doi.org/10.1145/3685527

Walker, C. P., & Timoneda, J. C. (n.d.). Identifying the sources of ideological bias in GPT models through linguistic variation in output.

Wang, C., Lee, B., Drucker, S., Marshall, D., & Gao, J. (2025, February). Data Formulator 2: Iterative Creation of Data Visualizations, with AI Transforming Data Along the Way [arXiv:2408.16119 [cs]]. https://doi.org/10.48550/arXiv.2408.16119

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-t., Tu, Z., & Lyu, M. (2024, August). Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6349–6384). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.345

Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., & Plank, B. (2024, August). "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.),

Findings of the Association for Computational Linguistics: ACL 2024 (pp. 7407–7416). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.441

Wankmüller, S. (2024). Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis [_eprint: https://doi.org/10.1177/00491241221134527]. *Sociological Methods & Research*, *53*(4), 1676–1752. https://doi.org/10.1177/00491241221134527

Webb, B. (2024, August). Synthetic Survey Respondents Creator. Retrieved April 8, 2025, from https://github.com/brockwebb/Synthetic-Survey-Respondents-Creator

Wuttke, A., Aßenmacher, M., Klamm, C., Lang, M. M., Würschinger, Q., & Kreuter, F. (2024, September). AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers [arXiv:2410.01824 [cs]]. Retrieved October 9, 2024, from http://arxiv.org/abs/2410.01824

Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions [arXiv:1905.10700 [cs]]. *ACM Transactions on Computer-Human Interaction*, *27*(3), 1–37. https://doi.org/10.1145/3381804

Zarouali, B., Araujo, T., Ohme, J., & de Vreese, C. (2023). Comparing Chatbots and Online Surveys for (Longitudinal) Data Collection: An Investigation of Response Characteristics, Data Quality, and User Evaluation. *Communication Methods and Measures*, 1–20. https://doi.org/10.1080/19312458.2022.2156489

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023, December). Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7915–7927). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.491

Zheng, M., Pei, J., Logeswaran, L., Lee, M., & Jurgens, D. (2024, November). When "A Helpful Assistant" Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 15126–15154). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.888

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can Large Language Models Transform Computational Social Science? [Place: Cambridge, MA Publisher: MIT Press]. *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Zou, Z., Mubin, O., Alnajjar, F., & Ali, L. (2024). A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires. *Scientific Reports*, *14*(1), 2781. https://doi.org/10.1038/s41598-024-53255-1

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

2.1 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 Jr. & Licato, 2023; Lee et al., 2023), conducting interviews (Chopra & Haaland, 2023; Cuevas 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; J. Kim & Lee, 2023), detecting non-human respondents in online surveys (Lebrun et al., 2024), 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., 2024; 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 U.S. 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 U.S. population is even more questionable than within the U.S., 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 U.S. 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.

2.2 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., J. 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, i.e., 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 U.S. 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, 2024). 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 U.S. 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. (2024) 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 U.S. 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 5% 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 U.S. 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., J. W. Kim et al., 2021; Muhlberger, 2003; Tucker et al., 2018), Internet discussions, however, often tend to conflate political complexities to two camps (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 Union, 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 and Shaw, 2013). Overall, the "knowledge sources" of LLMs are heavily concentrated on the English-speaking U.S. 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 compare to 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. (2024) 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. (2024) 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., Li et al., 2024; von der Heyde et al., 2024)[1], with prompting in non-English languages resulting in more U.S.-centric output nevertheless (e.g., Durmus et al., 2024; Havaldar et al., 2023; Johnson et al., 2022; W. Wang et al., 2024). 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

[1] Chapter 3 of this dissertation (reference refers to paper preprint).

left-of-center, environmentalist Green party (Greens), the Left party, and, more recently, the far-right "protest" party "Alternative for Germany" (AfD).[2] 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 under-represented 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., Brooks et al., 2006; Ford & Jennings, 2020; Lipset & Rokkan, 1967; Sass & Kuhnle, 2023). Moreover, the determinants of voting behavior on the micro-level play out in a different way than in the U.S. context: Partisanship and traditional socio-economic and religious cleavages and their impact on voting behavior have declined (Berglund et al., 2005; Dalton, 2014; Elff & Rossteutscher, 2011; Franklin et al., 2004; Jansen et al., 2013; Schmitt-Beck et al., 2022a, 2022b). At the same time, the socio-cultural dimension (Inglehart, 1977; Schmitt-Beck et al., 2022b) 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).

2.3 Data and Methods

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 1.1 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., 2024; 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, 2024), 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.

[2] 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.

Prompt creation

For the prompts provided to GPT-3.5, we create personas individually simulating each of the 1,905 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. Klein, 2014; Schmitt-Beck et al., 2022a, 2022b; Schoen et al., 2017). 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.[3] Missing values on any of the variables are imputed for 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 1.10). We then feed these personas as prompts to GPT-3.5 in German, using the completions-API,[4] 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 2.1, translated to English for illustrative purposes (see Appendix 1.2 for the German original).

*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 2.1: Example prompt (translated from German, variables **bolded** for emphasis).

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-U.S.- and/or -English context, in order to inform potential applications outside of the U.S. 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

[3] For details on the variables in the prompt, see Appendices 1.1 and 1.2. For summary statistics, see Appendix 1.6.
[4] For details, see Appendix 1.3.

Durmus et al., 2024). 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 present paper.

LLM configuration

Based on the outputs of a pilot test (see Appendix 1.3 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.[5] 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 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.[6] 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 (X. 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 (van Buuren, 2018). Specifically, we sample five completions per persona and estimate the variance between these samples.[7] 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 1.4), also considering non-voters and invalid votes. 1,427 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 1.5 for details and Appendix 1.10 for an investigation of systematic patterns in these personas/completions).

[5] See Appendix 1.3 for detailed explanations.

[6] Voting behavior is inherently situational. Therefore, *perfectly* capturing it with either survey or LLM data is unlikely. We discuss this in the Discussion section.

[7] 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.

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[8] 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, 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 (van Buuren, 2018). 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 $\sqrt{\mu(SE_\beta^2) + 1.2\sigma_\beta^2}$ (van Buuren, 2018) 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, 2024), *nnet* (Venables & Ripley, 2002), and *marginaleffects* (Arel-Bundock, 2021).

2.4 Results

Aggregate Performance

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

[8] Scores range from 0 to 1, with higher values indicating better predictive performance. For a detailed explanation, see Appendix 1.7.

Figure 2.2). For the two major parties, CDU/CSU and SPD, there are no significant differences.

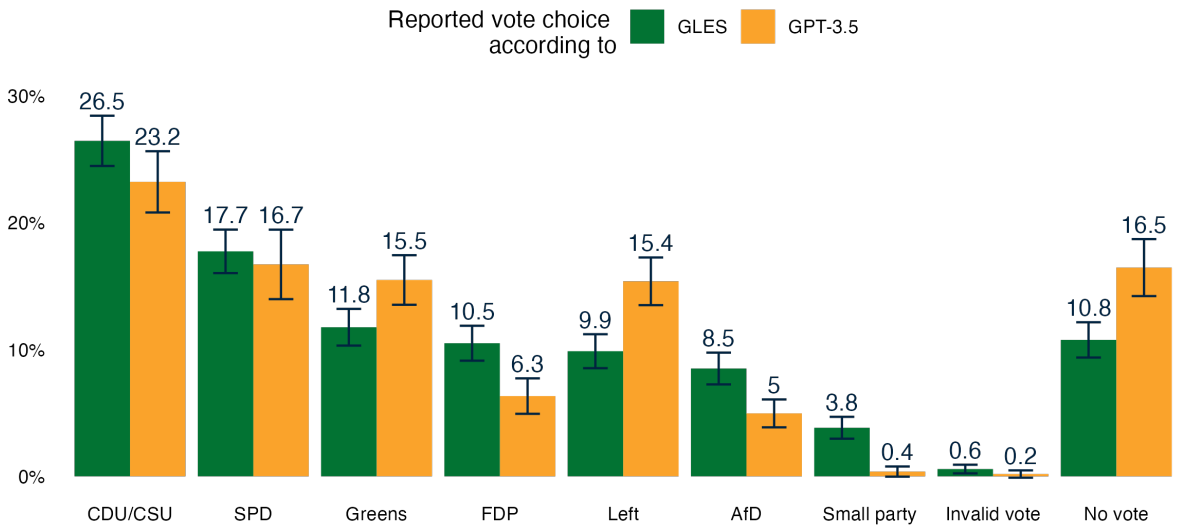


Figure 2.2: Distribution of vote shares as estimated by GLES and GPT.

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 1.7). 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 1.7). These differences are informative for both the aggregate and subgroup analyses. For the aggregate, they confirm what the overall distribution of 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.

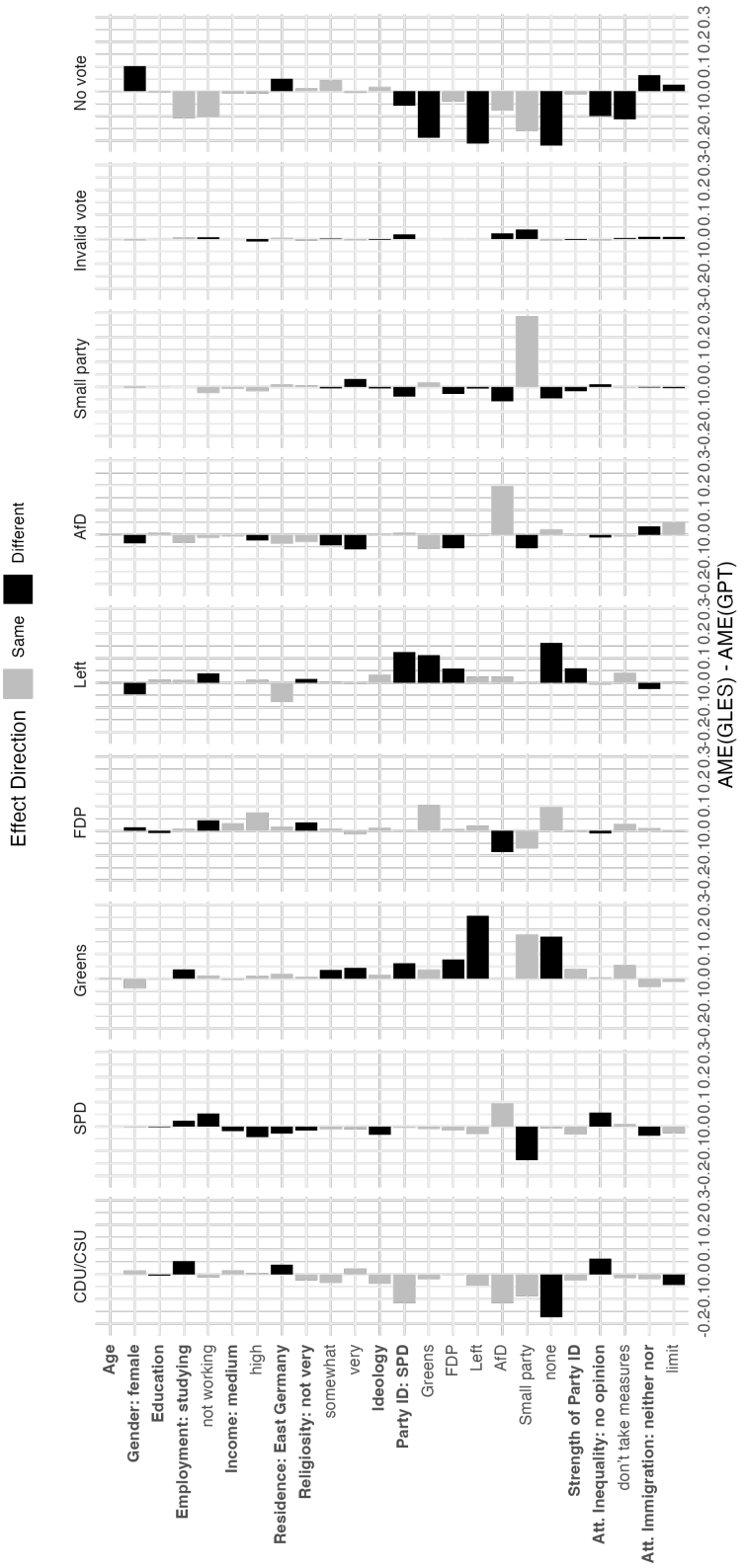


Figure 2.3: Difference in average marginal effects of prompt variables on vote choice, as estimated by GLES and GPT.

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 Appendix 1.8.

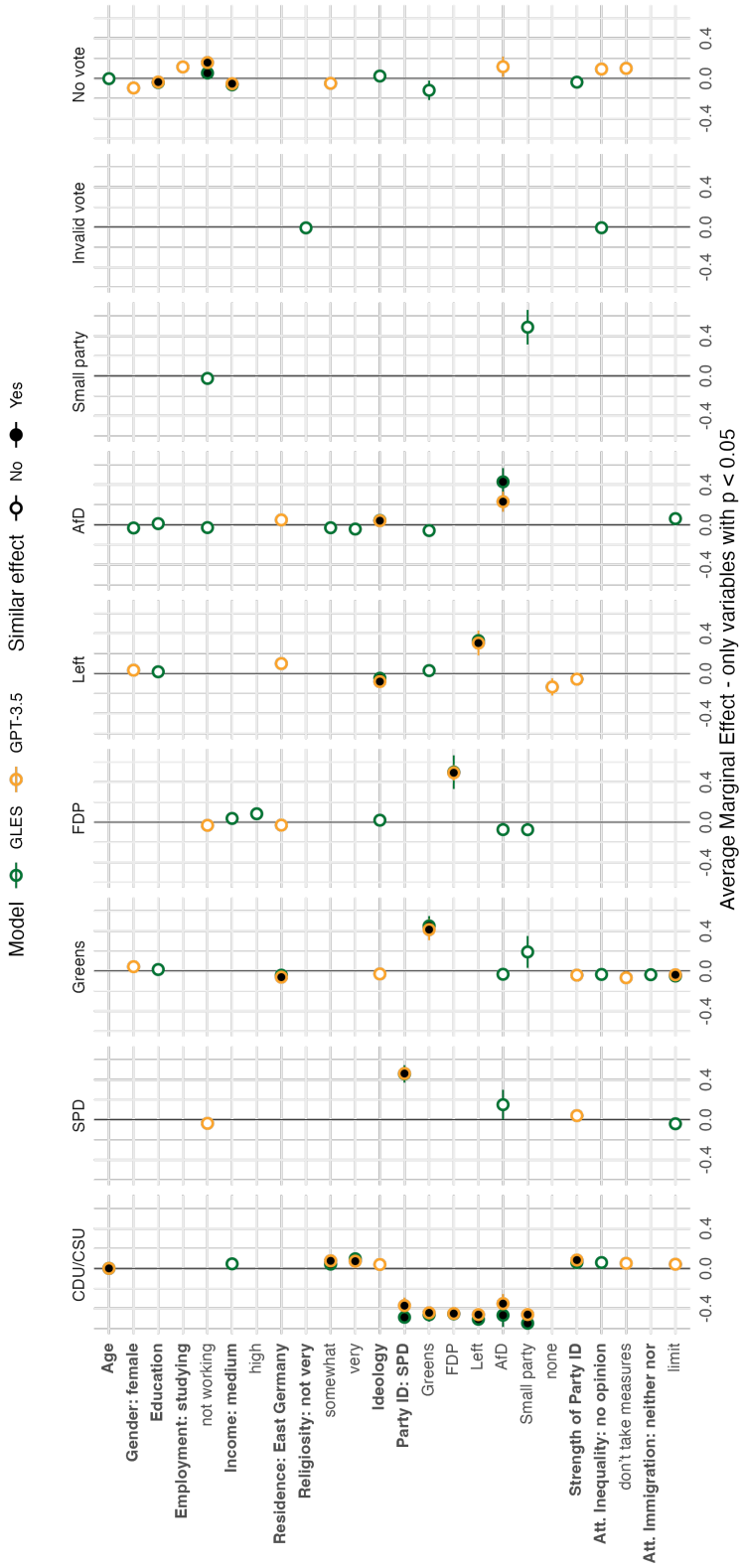


Figure 2.4: Average marginal effects of prompt variables on vote choice, as estimated by GLES and GPT. *Note: Average marginal effects describe the average of the fitted results of the model after first making individual predictions for each row in the original dataset, mirroring the real data (c.f. 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 Appendix 1.8.*

Subgroup Performance

Comparing the impact the prompting variables have on actual GPT- versus GLES-reported vote choice in multinomial regressions (see Figure 2.3 for a visualization of the differences in effect size accounting for effect direction, and Appendix 1.8 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 2.4, 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 Appendix 1.9), 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.

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

[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.

its predictions, without necessarily considering other, more complex mechanisms in the individual voting-decision making process.

2.5 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.

Regarding 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 U.S. Indeed, even studies considering the U.S. come to diverging conclusions (c.f. Argyle et al., 2023; Chu et al., 2023 vs. Bisbee et al., 2024; Dominguez-Olmedo et al., 2024; 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., 2024), and that this variance depends on the 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 LLMs. 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., 2024 to Ahnert et al., 2025; 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", i.e., 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 multi-level 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., 2025; Holtdirk et al., 2024; J. Kim & Lee, 2023) is a first step in this direction.

2.6 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.

References

Aaru. (n.d.). Aaru: Rethinking the Science of Prediction. Retrieved November 21, 2024, from https://aaruaaru.com

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2025). Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 15–36. https://doi.org/10.1609/icwsm.v19i1.35801

Arel-Bundock, V. (2021, September). Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests [Institution: Comprehensive R Archive Network Pages: 0.21.0]. https://doi.org/10.32614/CRAN.package.marginaleffects

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 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 [_eprint: 2406.09289]. https://arxiv.org/abs/2406.09289

Berglund, F., Holmberg, S., Schmitt, H., & Thomassen, J. (2005, July). 5 Party Identification and Party Choice. In *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. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 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*, 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., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 1877–1901, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

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, March). Language Models Trained on Media Diets Can Predict Public Opinion [arXiv:2303.16779 [cs]]. https://doi.org/10.48550/arXiv.2303.16779

Chua, G. (2024). An AI polling startup makes its predictions for the 2024 US election. *Semafor*. Retrieved November 21, 2024, from 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, February). Data Representativity for Machine Learning and AI Systems [arXiv:2203.04706 [cs, stat]]. Retrieved April 4, 2023, from http://arxiv.org/abs/2203.04706

Cuevas, A., Brown, E. M., Scurrell, J. V., Entenmann, J., & Daepp, M. I. G. (2023, October). Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models [arXiv:2309.10187 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2309.10187

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, September). *Political Realignment: Economics, Culture, and Electoral Change* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780198830986.001.0001

Delve AI. (n.d.). AI Persona Generator. Retrieved November 21, 2024, from https://www.delve.ai/ai-persona-generator

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2024). Questioning the Survey Responses of Large Language Models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information Processing Systems* (pp. 45850–45878, Vol. 37). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/515c62809e0a29729d7eec26e2916fc0-Paper-Conference.pdf

Durmus, E., Nguyen, K., Liao, T., 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. (2024). Towards Measuring the Representation of Subjective Global Opinions in Language Models. *First Conference on Language Modeling*. https://openreview.net/forum?id=zl16jLb91v

Elff, M., & Rossteutscher, S. (2011). Stability or Decline? Class, Religion and the Vote in Germany [Number: 1]. *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, April). *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). 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*. https://doi.org/10.1037/met0000540

Hartmann, J., Schwenzow, J., & Witte, M. (2023, January). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation [arXiv:2301.01768 [cs]]. Retrieved March 28, 2023, from http://arxiv.org/abs/2301.01768

Havaldar, S., Singhal, B., Rai, S., Liu, L., Guntuku, S. C., & Ungar, L. (2023, July). Multilingual Language Models are not Multicultural: A Case Study in Emotion. In J. Barnes, O. De Clercq, & R. Klinger (Eds.), *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 202–214). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wassa-1.19

Heiss, A. (2022, May). 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. Retrieved July 10, 2024, from 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*, peps.12543. 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 (A. Sánchez, Ed.) [Number: 6]. *PLoS ONE*, *8*(6), e65782. https://doi.org/10.1371/journal.pone.0065782

Holtdirk, T., Assenmacher, D., Bleier, A., & Wagner, C. (2024, October). 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* [OCLC: 979580560]. Princeton University Press.

International Telecommunication Union. (2022). *Measuring digital development Facts and Figures 2022* (tech. rep.).

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 [Number: 2]. *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, March). The Ghost in the Machine has an American accent: Value conflict in GPT-3 [arXiv:2203.07785 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2203.07785

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

Kim, J., & Lee, B. (2023, November). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction [arXiv:2305.09620 [cs]]. Retrieved January 23, 2024, from http://arxiv.org/abs/2305.09620

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. (2024, May). Rgpt3: Making requests from R to the GPT API. https://doi.org/10.5281/zenodo.7327667

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*. 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 Jr., A., & Licato, J. (2023, July). Generating Better Items for Cognitive Assessments Using Large Language Models. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 414–428). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.34

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. (2024). Detecting the corruption of online questionnaires by artificial intelligence. *Frontiers in Robotics and AI, Volume 10 - 2023*. https://doi.org/10.3389/frobt.2023.1277635

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

Liew, A., & Mueller, K. (2022, December). Using Large Language Models to Generate Engaging Captions for Data Visualizations [arXiv:2212.14047 [cs]]. Retrieved October 17, 2023, from http://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, September). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve [arXiv:2309.13638 [cs]]. Retrieved September 18, 2024, from http://arxiv.org/abs/2309.13638

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, *11*(1). https://doi.org/10.1177/20531680241231468

Mendoza, D. (2024). AI polling company defends wrong predictions on the US election. *Semafor*. Retrieved November 21, 2024, from 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*. 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). https://doi.org/10.1515/comm.2003.009

Nicoletti, L., & Bass, D. (2023). Humans Are Biased. Generative AI Is Even Worse. *Bloomberg.com*. Retrieved November 24, 2023, from https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Nie, E., Shao, B., Ding, Z., Wang, M., Schmid, H., & Schütze, H. (2024, June). BMIKE-53: Investigating Cross-Lingual Knowledge Editing with In-Context Learning [arXiv:2406.17764 [cs]]. Retrieved July 11, 2024, from http://arxiv.org/abs/2406.17764

Nie, E., Yuan, S., Ma, B., Schmid, H., Färber, M., Kreuter, F., & Schütze, H. (2024, February). Decomposed Prompting: Unveiling Multilingual Linguistic Structure Knowledge in English-Centric Large Language Models [arXiv:2402.18397 [cs]]. Retrieved July 11, 2024, from http://arxiv.org/abs/2402.18397

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

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 Number: 6]. https://doi.org/10.48550/ARXIV.2303.08774

R Core Team. (2024, July). R: The R Project for Statistical Computing. Retrieved July 2, 2024, from https://www.r-project.org/

Roth, M. (2024). Ask a Llama - Creating variance in synthetic survey data. General Online Research. https://doi.org/https://doi.org/10.17605/OSF.IO/CHBTR

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023, June). Towards Coding Social Science Datasets with Language Models [arXiv:2306.02177 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2306.02177

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023, July). Whose Opinions Do Language Models Reflect? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 29971–30004, Vol. 202). PMLR. https://proceedings.mlr.press/v202/santurkar23a.html

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. (2022a, April). The Changing German Voter. In R. Schmitt-Beck, S. Roßteutscher, H. Schoen, B. Weßels, & C. Wolf (Eds.), *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, April). 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* (First edition). Oxford University Press.

Schumacher, S., & Kent, N. (2020, April). 8 charts on internet use around the world as countries grapple with COVID-19. Retrieved October 17, 2023, from 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 [Bandiera_abtest: a Cg_type: World View Publisher: Nature Publishing Group Subject_term: Ethics, Machine learning, Technology, Scientific community]. *Nature*, *616*(7957), 413–413. 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 [Publisher: Routledge _eprint: https://doi.org/10.1080/13645579.2020.1844896]. *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. *2023 IEEE Visualization and Visual Analytics (VIS)*, 231–235. https://doi.org/10.1109/VIS54172.2023.00055

Synthetic Users. (n.d.). Synthetic Users. Retrieved November 21, 2024, from 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 Number: 1]. 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). How AI reduces the world to stereotypes. Retrieved November 24, 2023, from https://restofworld.org/2023/ai-image-stereotypes/

van Buuren, S. (2018, July). *Flexible Imputation of Missing Data, Second Edition* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9780429492259

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R [Number: 3]. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

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

von der Heyde, L., Haensch, A.-C., Wenz, A., & Ma, B. (2024). United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections [Version Number: 2]. https://doi.org/10.48550/ARXIV.2409.09045

W3Techs. (2024, July). Usage Statistics and Market Share of Content Languages for Websites, July 2024. Retrieved July 2, 2024, from https://w3techs.com/technologies/overview/content_language

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-t., Tu, Z., & Lyu, M. (2024, August). Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6349–6384). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.345

Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., & Plank, B. (2024, August). "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.),

Findings of the Association for Computational Linguistics: ACL 2024 (pp. 7407–7416). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.441

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 [Number: 43]. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wikipedia. (2024, July). Wikipedia:Wikipedians - Wikipedia. Retrieved July 2, 2024, from 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 [Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2020.1785067]. *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* [OCLC: 819323088]. Cambridge University Press.

3 United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections

3.1 Introduction

Large language models (LLMs) have recently emerged as a new tool for researchers in computational social science and have been proposed to complement existing methods for understanding human attitudes and behaviors. For example, research has started to assess to what extent LLM-generated "synthetic samples" can be used as a viable and efficient alternative for collecting data about public opinion (Argyle et al., 2023; Bisbee et al., 2024; Dominguez-Olmedo et al., 2024; Sanders et al., 2023). Textual LLM output reflects a probability of how likely a given word is followed by another word, conditional on the training data and contextual information provided in the specific prompt. Since LLMs are trained on large amounts of human-generated text data, their output has been argued to reflect human attitudes and behaviors. Thus, researchers hope that by conditioning an LLM with specific individual-level information in the input, the LLM could be prompted to respond from that individual's perspective and – if scaled up – synthesize public opinion data for entire human samples. Such synthetic samples have been proposed to ease the collection of previously unobserved public opinion data, including, but not limited to, data about hard-to-survey populations, sensitive topics, or future outcomes, and allow for fast and low-cost questionnaire pre-testing and pilot studies. As a result, an increasing number of enterprises offer survey and market research "solutions" based on LLM-synthetic samples (e.g., Aaru, n.d. Delve AI, n.d. Levanti & Verret, 2024; Synthetic Users, n.d.). For example, AI startups have tried (and failed) to use synthetic samples for election predictions (Chua, 2024, but see S. Jiang et al., 2024). Elections are inherently challenging to predict due to the complexity of human voting behavior. This implies that it is unlikely for LLMs to perform any better than the plethora of existing methods and data that have struggled, especially given that LLMs rely on previous human knowledge. Yet, the surge of startups capitalizing on this challenge and the general hope (and hype) that is being put in AI by selling AI "snake oil" (e.g., Mendoza, 2024) persists, calling for systematic research about biases in LLM-based predictions to inform science and industry.

Previous research has found biases in LLM output with regard to various outcomes, including political attitudes and psychological measures (Atari et al., 2023; Durmus et al., 2024; Kim & Lee, 2023; Lee et al., 2023; Sanders et al., 2023; Santurkar et al., 2023; von der Heyde et al., 2025; P. Wang et al., 2024)[1]. Among the potential reasons for why these biases occur, unrepresentative training data regarding linguistic, social, political, and digital contexts are often mentioned (Bender et al., 2021; Cao et al., 2023; von der Heyde et al., 2025). With more than

[1] Chapter 2 of this dissertation.

50% of Internet content estimated to be English, the amount of available native-language training data for LLMs is considerably smaller for countries with any other native language. Moreover, the relationship between societal and political structure and public opinion formation differs between countries, and is likely not sufficiently represented in LLM training data. Finally, the training data is likely affected by coverage bias caused by the "digital divide". There may be differences between the respective target population and those who contributed to the specific texts selected for training LLMs. While it is difficult to identify their exact causes due to LLMs being "black boxes", such biases can challenge the validity of findings based on LLM-synthetic samples, and risk reinforcing existing biases in social science research, policymaking, and society. Therefore, computational social scientists need to investigate if and under which conditions LLM-generated synthetic samples can be applied for public opinion prediction by comparing different linguistic, political, social, and digital contexts.

Initial studies that used LLM-synthetic samples for estimating public opinion, particularly vote choice, yielded results matching survey data in the context of the U.S. general population (Argyle et al., 2023; Kim & Lee, 2023; Lee et al., 2023). More recent research, however, has challenged these initial findings, particularly in other national contexts (Durmus et al., 2024; Motoki et al., 2023; Qu & Wang, 2024) and languages (Qi et al., 2024; von der Heyde et al., 2025). These studies find evidence for politically left-leaning, and culturally and linguistically WEIRD (Western, Educated, Industrialized, Rich, Democratic) biases in LLM outputs, reproducing simplified stereotypes for other national and linguistic contexts. However, existing cross-national studies on attitudinal biases in LLMs have typically employed *country-level* prompting only (Atari et al., 2023; Durmus et al., 2024; Motoki et al., 2023), not individual-level personas sourced from survey data. Such generic, country-level input only allows for generic output, not testing LLMs' capabilities in producing estimates of public opinion based on nuanced, individual-level predictions which might potentially result in better aggregate results. Previous research that used LLM-synthetic samples based on *individual-level* characteristics for estimating public opinion, in turn, has identified biases regarding certain subgroups, but, thus far, has mostly been conducted in isolated national settings, not making any cross-national comparisons (Kim & Lee, 2023; Lee et al., 2023; Sanders et al., 2023; Santurkar et al., 2023; von der Heyde et al., 2025; A. Wang et al., 2025). An exception is the study by Bisbee et al. (2024), which finds in a supplementary cross-national analysis that ChatGPT performs similarly poorly across countries, with a tendency to predict attitudes that are more common in the benchmark survey data. Although the authors argue that poor performance can be expected for non-native English speaking contexts, the U.S. accuracy scores are, in fact, among the lowest. The study does not test whether performance is related to the prompt language (they only used English) or content (although they found effects of prompt wording in their main, U.S.-focused study).

Additionally, while most existing studies employed LLM-synthetic samples for "predicting the past" (Argyle et al., 2023; Bisbee et al., 2024; Durmus et al., 2024; Qi et al., 2024; Qu and Wang, 2024, but see S. Jiang et al., 2024; Kim and Lee, 2023), it is essential to assess their performance in making predictions of unobserved outcomes, such as future election results. Such an investigation can illustrate how past training and survey data informs the prediction of future outcomes and how much past information is necessary for accurate prediction. Indeed, if detailed and timely individual-level survey data is necessary to make somewhat accurate predictions with LLM-generated samples, such samples may not be of much use to researchers, as they would still need to resort to surveys and possibly could even ask about the unobserved outcome of interest directly in those surveys.

In this paper, we aim to bridge these gaps between cross-national, individual-level, and

future-outcome investigations and applications of LLM-synthetic samples. Using the example of LLM-based predictions of voting behavior in the 2024 European Parliament elections, we examine to what extent LLM-based predictions of individual public opinion exhibit context-dependent biases by addressing the following research questions:

RQ1. Can LLMs predict the aggregate results of future elections?
RQ2. How does LLMs' predictive performance differ across countries?
RQ3. How does LLMs' predictive performance differ across prompt languages?
RQ4. How does LLMs' predictive performance differ depending on the information provided in the prompt?

Elections are a real-world example of an important, yet challenging prediction task in public opinion research. More specifically, the 2024 European Parliament elections provide a relevant test case for biases in LLM-based predictions across contexts, featuring both a comparable temporal and electoral reference point and high diversity in linguistic, social, political, and digital contexts across the 27 European Union (EU) member states. For an additional in-depth investigation of differences in LLMs' bias across languages, we select six countries (France, Germany, Ireland, Poland, Slovakia, and Sweden), differing in native language Internet coverage, linguistic prevalence within the EU, language family, as well as in population size, geographic and political position within, and attitudinal position towards Europe. Following previous research (Argyle et al., 2023; Bisbee et al., 2024; Dominguez-Olmedo et al., 2024), we employ the synthetic sampling approach: We sequentially prompt the LLM GPT-4-Turbo (OpenAI et al., 2023) with pseudonymized individual-level background information from an existing survey sample of approximately 26,000 eligible voters – the Eurobarometer 99.4 from summer 2023 (European Commission, 2024). For each individual, we create a description including socio-demographic and attitudinal information to prompt the LLM. Prompts vary with regards to citizens' age, gender, education, socio-economic class, occupation status, and urbanicity. Additionally, the profiles include information about individuals' political interest, ideology, trust in the EU, and attitude towards European integration, as well as the parties competing in the respective country. Before the European elections have taken place, we then ask the LLM to predict each person's voting behavior. As a robustness check, we perform the same analyses on two open-source LLMs, Llama 3.1 (Dubey et al., 2024) and Mistral (A. Q. Jiang et al., 2023), using the same prompts and model configuration for input.

Applying the synthetic sampling approach to election prediction, we show (1) how well popular LLMs perform at predicting *future* voting behavior based on past training and *individual-level* survey data, (2) how this performance differs *across national* and *linguistic* contexts, and (3) whether it is currently feasible to supplement survey data with LLM-based data given limited individual-level *information* provided in the prompt. In investigating the contextual differences of LLM-based predictions of public opinion, our research contributes to the understanding and mitigation of biases and inequalities in the development of LLMs and their applications in computational social science.

3.2 Data and Methods

Sample and LLM selection

To examine biases of LLM-synthetic samples in a variety of linguistic, social, political, and digital contexts, our study spans all 27 EU member states (EU-27). For an additional in-depth investigation of differences in LLMs' predictive performance across languages, we select five countries differing in native language Internet coverage (W3Techs, 2024), linguistic prevalence within the EU, language family, as well as in population size, geographic and political position within, and attitudinal position towards Europe (for details, see Appendix 2.3): France, Germany, Poland, Slovakia, Sweden, and Ireland (as an English-language baseline).

To create a realistic sample of individual-level profiles on which we base our predictions of vote choice in the 2024 European Parliament elections, we rely on the most recent available Eurobarometer survey data (EB 99.4) from May-June 2023. This data has been collected with face-to-face interviews of EU citizens aged 15 years and over and resident in the EU-27, based on stratified, multi-stage probability samples (European Commission, 2024). From this data, only voting-eligible EU citizens are selected, resulting in a sample of about n=1000 per EU member state (with the exception of Luxembourg and Malta, with a sample size of about n=500 each) or about 26,000 respondents in total. For summary statistics of all variables, see Appendix 2.2.

Simulating a realistic use-case, we use one of the most popular and powerful LLMs at the time of conducting the study, GPT-4-Turbo (version 2024-04-09). This model has the most recent training data corpus of all GPT models, with a cutoff date in December 2023 (OpenAI, n.d.). Further, it is supposed to have better multilingual capacities, be better at solving complex instructions, and less likely to "hallucinate", that is, provide fabricated output. Finally, its performance in predicting public opinion was shown to be better when adding information beyond demographics (Lee et al., 2023), and in different languages (W. Wang et al., 2024). To understand whether any biases we find can be generalized across LLMs, and to guide the development of future LLMs, we perform the same analyses on two open-source LLMs, Llama 3.1 (Knowledge cutoff December 2023 Dubey et al., 2024) and Mistral (Knowledge cutoff December 2022, A. Q. Jiang et al., 2023). Llama is optimized for multilingual dialogue use cases and supposed to be comparable to GPT but superior to other open source LLMs. Mistral is chosen for its robust performance in handling a wide range of tasks, including those requiring content reasoning and creative writing, which was shown to complement and in some cases even surpass the strengths of Llama models (MistralAI, 2023).

Prompt creation

For each individual in the Eurobarometer sample, we create a description including sociodemographic and attitudinal information with which we prompt the LLMs using second-person pronouns (Bisbee et al., 2024). Prompts vary with regards to citizens' age, gender, education, socio-economic class, occupation status, and urbanity. Additionally, the profiles include information about individuals' political interest, ideology, trust in the EU, and attitude towards European integration. These variables have been identified as determinants of voting behavior in EU elections (Braun & Schäfer, 2022; Ford & Jennings, 2020; Giebler & Wagner, 2015). In using vote choice as a test case for examining biases in predictions based on LLM-synthetic samples, we simulate a realistic use case where researchers and practitioners with limited resources and *limited*

information about their target population rely on *off-the-shelf* LLMs. The aim of the study is to assess the quality and *systematic differences* of LLMs' predictions across contexts when *holding information constant*. Thus, we do not account for country-specific determinants of voting behavior in European elections to ensure cross-country comparability of the prompts, even though this might limit the LLMs' predictive accuracy. Finally, the prompts feature the parties competing in the respective country that a) currently have a seat in the European Parliament or b) polled above the respective country's electoral threshold at the time of data collection (for details on which parties these are, see Appendix 2.1). For countries that do not have an electoral threshold in EU elections, we require a minimum of 2%, as this is the minimum threshold all countries have to implement by the 2029 EU elections (Sabbati & Grosek, 2023). We also include the pan-European party Volt for countries in which it is competing. The existence of parties below this threshold is indicated by "several smaller parties" at the end of the list. The order of parties is randomized to avoid any tendencies by GPT models to choose (one of) the first option(s) listed (Brand et al., 2023; Pezeshkpour & Hruschka, 2024). However, the order of variables, that is, sentences in the prompt, is not randomized, as this would risk contrived language that could impact the predictive accuracy and because assessing the impact of prompt engineering on prediction quality is not the aim of this study. Missing values on any of the individual variables are imputed for n = 6800 respondents (26% of the total sample) using multivariate imputation by chained equations (van Buuren & Groothuis-Oudshoorn, 2011). Clarifying the year and that the individual is voting-eligible aims to avoid erroneous predictions due to assumed ineligibility (von der Heyde et al., 2025). In order to avoid refusals by the LLMs and therefore incomplete output, we include a request to make a prediction solely on the data provided. Finally, the instruction to keep the answer short aims to ensure cost-efficiency, as the LLMs might otherwise unnecessarily elaborate on their choices. Figure 3.1 shows an exemplary prompt. For details on the prompting variables and their Eurobarometer source, see Appendix 2.1.

For the entire sample spanning all EU member states, we prompt the LLMs with the individual profiles in English, once containing only socio-demographic information, and once also containing the attitudinal variables (underlined in Figure 3.1). Additionally, for profiles from the five selected countries, we prompt the LLMs in the respective country's native language. For translations of the prompts, performed by researchers native in the respective language and on the basis of the native version of the questionnaire, see Appendix 2.1.

> *The year is 2024. You are a voting-eligible citizen of an EU-member state living in* ***Germany****. The parties competing in the 2024 elections to the European Parliament are* ***CDU/CSU, SPD, Bündnis 90/Die Grünen, FDP, Die Linke, AfD, Volt, Freie Wähler, Tierschutzpartei, Familienpartei, Bündnis Deutschland, Bündnis Sahra Wagenknecht, Piratenpartei****, and several smaller parties.*
> *You are* ***29*** *years old. You are* ***female****. You have a* ***university*** *education. Economically, you are* ***upper middle*** *class. You are* ***working****. You live in a* ***big city****. You are* ***very*** *interested in politics. Ideologically, you are* ***center-left****. You* ***think*** *that more decisions should be taken at the EU level. You tend* ***to trust*** *the European Union.*
> *Will you vote in the 2024 elections to the European Parliament, and if so, for which party? Use the information above to make as good a prediction as possible, and keep your answer as short as possible, if possible just "No" or the name of the party you would predict.*

Figure 3.1: Example prompt (prompt variables in **bold**; attitudinal information underlined).

LLM configuration

We automate the data collection through the Azure OpenAI REST API for the GPT-based data (OpenAI et al., 2023), and through local instances of Llama and Mistral (Meta, n.d. MistralAI, n.d.). Azure OpenAI provides private, local instances of OpenAI's GPT models, thereby ensuring the input data is not passed on to third parties (i.e., OpenAI servers). Open-source LLMs, in contrast, typically can be downloaded and run on local computing infrastructure to begin with, minimizing privacy concerns. We employ zero-shot prompting and, in line with previous studies (Aher et al., 2023; Bisbee et al., 2024; Lee et al., 2023; Tjuatja et al., 2024), configure the LLMs to a temperature of 0.9. To further control the LLMs' completions' length, we limit the output to a maximum of 40 tokens. Having tested exemplary completions in all target languages, 40 tokens allow for a response including a complete sentence with all necessary information. As previous research showed little variance in individual vote choice predictions when prompting GPT repeatedly (von der Heyde et al., 2025), we only prompt the LLMs once per individual. We collect the data shortly before the European elections are held (between June 6 and 9, 2024, depending on the member state), between May 29 and June 4, 2024. Data for the robustness checks was collected on July 29 (Llama) and August 1 (Mistral) – however, as their knowledge cutoffs are before the elections took place (Dubey et al., 2024), this should not impact the results.

Vote choice extraction

Vote choices are extracted from LLM completions based on a set of predefined keywords per competing party, as well as non-voting and invalid voting (see Appendix 2.1). As European elections typically feature a large number of very small political parties beyond the ones established in national politics, votes for parties that do not meet the respective country's electoral threshold in the official results (Sabbati & Grosek, 2023), or, in cases of no threshold, parties that do not obtain a seat in the newly elected parliament (European Parliament, n.d.), are summarized as "Other" for the analyses that follow. As this study aims to depict a realistic use-case as opposed to optimizing predictions a priori, completions that do not contain a definite party choice are recorded as missing and only counted for turnout calculation if the prediction clearly states the person would have voted, but not for vote share calculations (see Appendix 2.2 for proportions of missing values).

Since we prompt the LLMs to keep their responses concise, we expect token probabilities to not differ much from the displayed text output – that is, we expect the displayed output to mostly correspond to the token with the highest probability. Therefore, we do not use token probabilities for analytical purposes, but rather examine the actual text output. Instruction-tuned models like GPT-4-Turbo have the advantage of making the text output directly accessible to users. We consider this to be the most straightforward approach we would expect users of LLM-synthetic samples to apply.

Analysis

We weight the extracted results with the Eurobarometer-provided weights to better approximate the target population. To answer our first research question, we compare the aggregate predicted voting behavior when prompted in English to the official national-level results across all 27 countries, differentiating between turnout and party vote shares among voters. Specifically,

we compare the mean and variance of predicted and actual turnout, as well as several metrics for correct party vote share prediction, including prediction of the winning party, the rank ordering of parties, and average absolute differences in party vote shares per country as well as across European parliamentary groups as announced in the post-election constitutive session (European Parliament, n.d.). We do not account for the different electoral systems in place in the different countries, nor for the different electoral thresholds, both of which impact voting behavior and vote aggregation, thereby potentially limiting the predictive accuracy. Future research could investigate whether such adjustments improve predictions.

To tackle our second research question, we contrast the differences in predicted turnout and party vote shares within the EU-27. We compare countries according to whether they have compulsory voting, their European region (EuroVoc, n.d.), and their language family (Wikipedia, 2024). In line with our third research question, we also analyze the LLMs' predictive performance based on prompts in English and the five selected countries' native languages. Regarding our fourth research question, we compare whether predictions containing the full set of information in the prompt perform better than those based solely on socio-demographic information.

In our analyses, we do not report confidence intervals or conduct traditional tests for statistical inference. Doing so would imply that the primary source of error stems solely from the sampling of observations – as is typically assumed in traditional survey research and related studies – even when LLM outputs are prompted using information derived from a sample. Instead, we argue that additional sources of error can arise from biases inherent in LLMs' data-generating process.

Data collection and analysis is conducted using the software R, version 4.3.2 (R Core Team, 2024), especially the packages *tidyverse* (Wickham et al., 2019), *mice* (van Buuren & Groothuis-Oudshoorn, 2011), *rgpt3* (Kleinberg, 2024), and *survey* (Lumley, 2004).

3.3 Results

Overall prediction of EU election results

Despite the capabilities of GPT-4-Turbo, we are still far from being able to use it as an accurate and reliable prediction tool for public opinion: Predictions of turnout and party vote shares in the 2024 European elections based on synthetic samples of the voting population fail. With an average predicted turnout of 83%, predictions based on GPT-4-Turbo overestimate turnout by 34 percentage points on average, not capturing the substantial variation between countries (Figure A1). Predictions of turnout almost all range above the total range of actual turnout. Considering party vote shares, GPT-4-Turbo-based predictions mostly fail to predict the winner (11 out of 27) or ranking of parties (Figure 3.2), with the LLM only identifying 8% of party ranks correctly on average (with a median of 0%). Predictions of individual party vote shares often differ greatly from the actual result (Figure 3.3), with average differences of five to nearly 17 percentage points per country. This average per country masks a high variation between parties, with larger differences between predicted and actual vote shares especially for parties not belonging to the Green or Left parliamentary groups (Figure A2.2), confirming findings from previous research and including the two biggest groups left and right of center and the newly formed right-wing groups. The former have suffered substantial national losses in recent years, which may not have been picked up by GPT-4-Turbo due to the temporal limits of its training data.

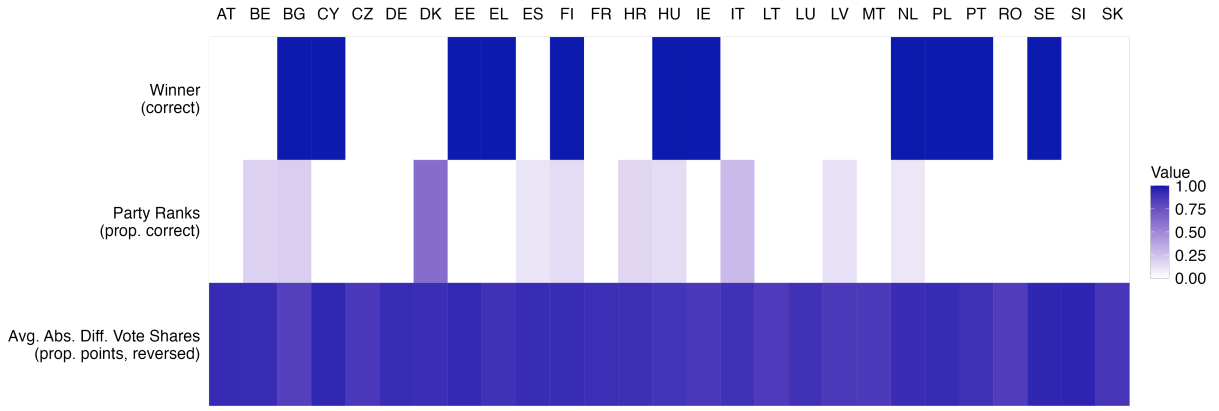


Figure 3.2: Predictive performance of GPT-4-Turbo for the 2024 EU election party results (based on full English prompt).

Note: Average absolute differences in vote shares have been reversed so that higher values correspond to better predictive performance in line with the other metrics. Example: an average absolute difference of 5 percentage points (0.05) would be displayed as 0.95.

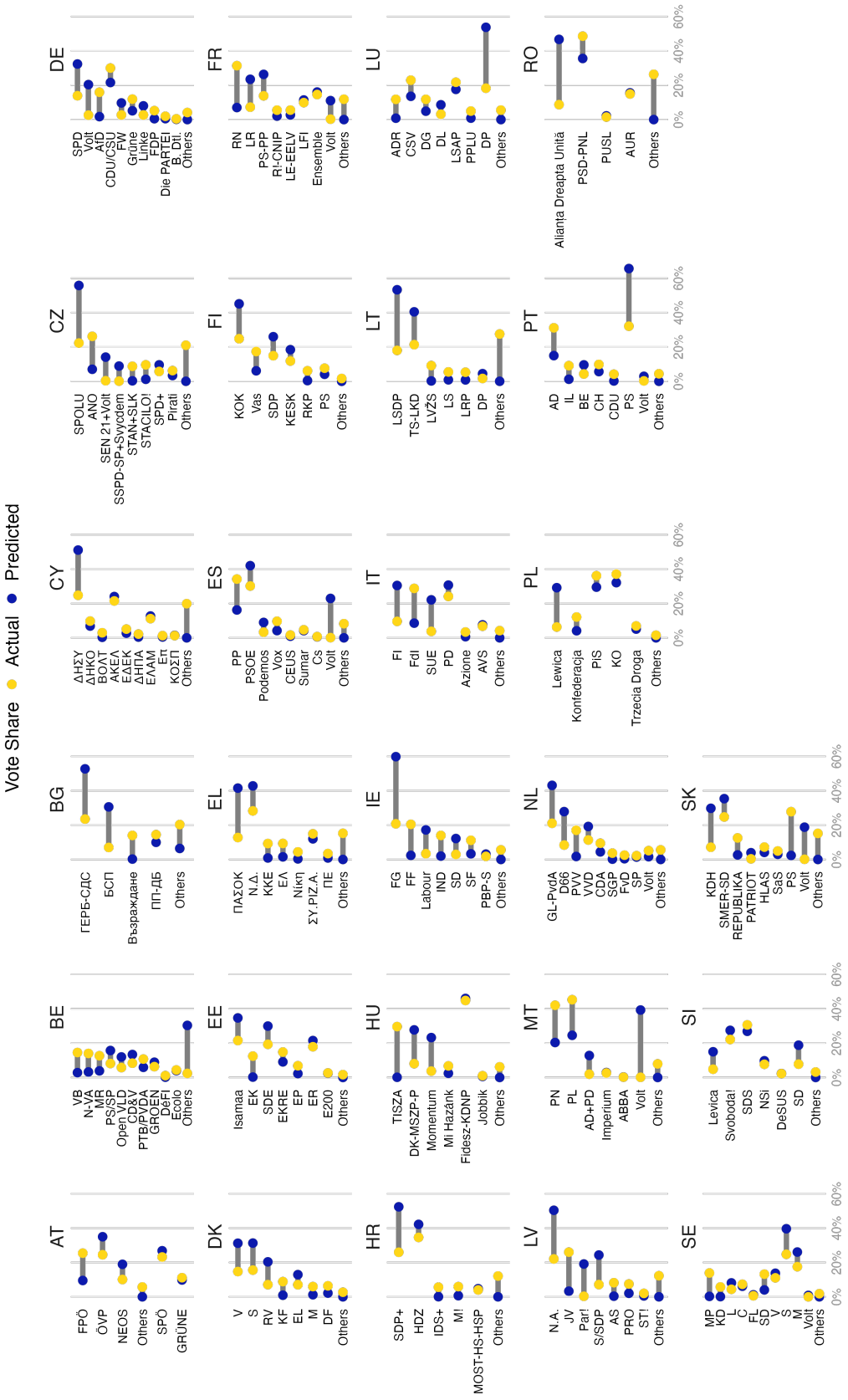


Figure 3.3: Differences between actual and predicted party vote shares by country and party (based on full English prompt).

Differences in predictive performance across countries

For English prompting overall, GPT-4-Turbo's predictive performance of turnout is higher for countries with high actual turnout (Figure 3.4), while it overestimates turnout especially for countries with typically low actual turnout. This pattern can be explained by the LLM's tendency to predict rather high turnout regardless of country, and holds even for the four countries with compulsory voting. The difference between predicted and actual turnout is among the lowest for Belgium and Luxembourg, Western European countries with French as an official language, one of the most dominant languages in Europe. In contrast, for Greece and Bulgaria, which are situated in South-East Europe and whose native languages use Cyrillic alphabets and are less commonly used, the differences are among the highest.

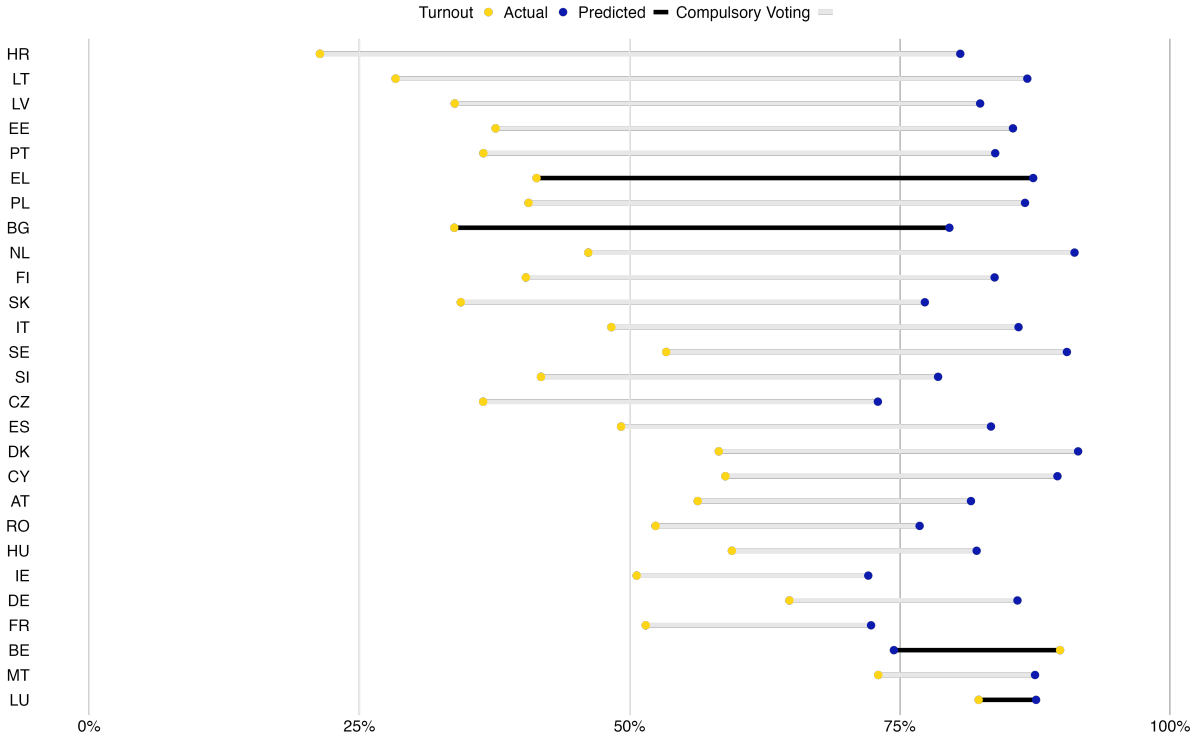


Figure 3.4: Difference between actual turnout in the 2024 EU elections and GPT-4-Turbo's predictions (based on full English prompt).

When differentiating regions, the LLM's overestimations of turnout tend to be higher for Eastern and Southern European countries, especially if considering the Baltic states (Lithuania, Latvia, and Estonia) as (historically) Eastern rather than (aspirationally) Northern European (Figure 3.5a). This pattern is confirmed when investigating native language families (Figure 3.5b): Overestimations of turnout are higher for Baltic and Slavic language countries when prompting GPT-4-Turbo in English. As especially Slavic languages are native to Eastern European countries, it is no surprise that these patterns overlap. Historically, turnout tends to be lower in non-Western European countries. However, the LLM is unable to capture this pattern, but assumes the high turnout levels of Western European countries that typically speak one of the more dominant Germanic or Romance languages, such as English, German, or French.

The same holds when it comes to predictions of party vote shares, which on average differ more from the actual results for Eastern and Southern European countries (again, especially when considering the Baltics as part of this group; Figure 3.5c/d) and such with Slavic or Baltic native languages. As evidenced by the case of Romania, a country with a Romance native language but among the countries with the highest difference between predicted and actual party vote shares, linguistic and geographical factors likely interact when it comes to GPT's predictive accuracy.

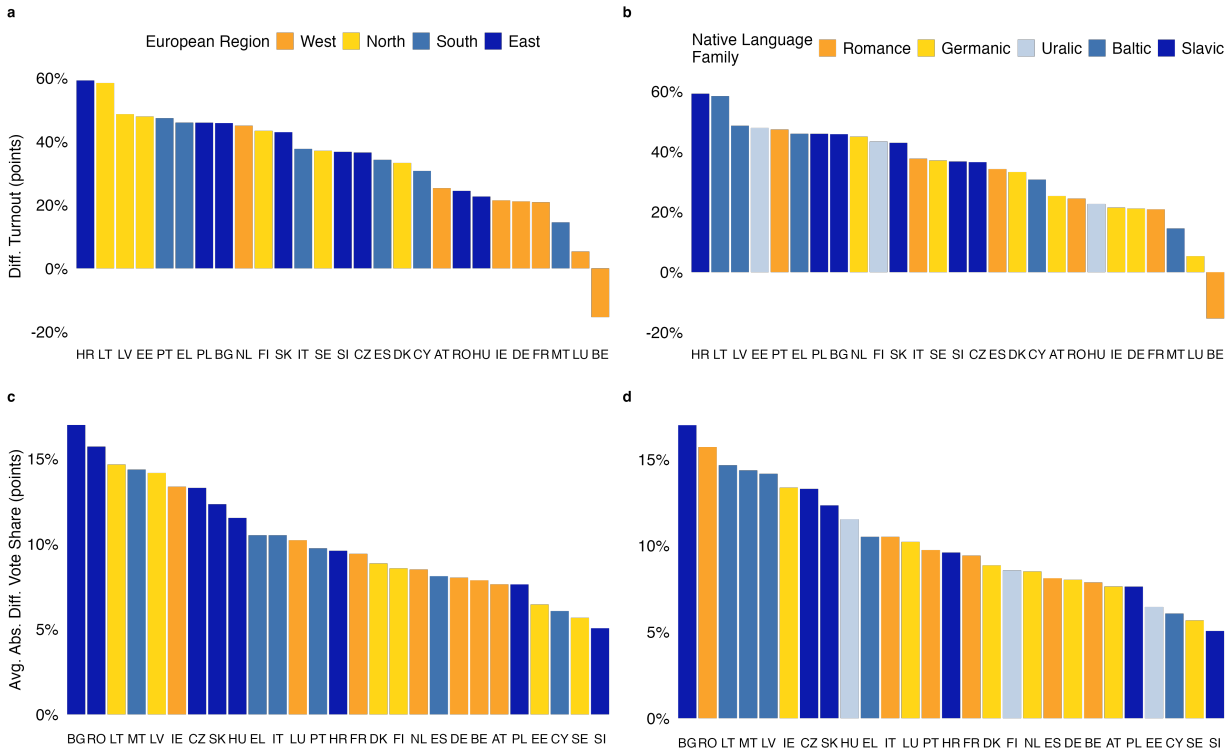


Figure 3.5: (Average) difference between actual turnout and party vote shares in the 2024 EU elections and GPT-4-Turbo's predictions (based on full English prompt) **by region and language family**.

Differences in predictive performance across prompt languages

In all of the five countries examined, prompting in the native language leads to an even bigger overestimation of turnout than when prompting in English. The difference in difference of turnout estimation between English and native-language prompting is especially strong for France, followed by Slovakia. While there is barely a difference between prompt languages for Poland, the overestimation is particularly large regardless of language, at over 40 percentage points. When it comes to differences in party vote shares, the pattern somewhat reverses (Figure 3.6b). Here, native-language prompting tends to outperform English-language prompting, at least in Germany and Sweden. For France and Poland, differences between prompt languages are not very large. Notably, the average difference between predicted and actual vote shares is highest for the benchmark Ireland. This may be due to Ireland's complex single-transferable (ranked choice) voting system, which is not accounted for by the LLM or the aggregation.

To summarize, English-language prompting returns better predictions than native-language

prompting for turnout (Figure 3.6a), but not as much for party vote shares (Figure 3.6b). The size of differences between English- and native-language prompting depends on the country in question, suggesting that GPT-4-Turbo is worse at predicting Eastern European voting behavior regardless of prompt language.

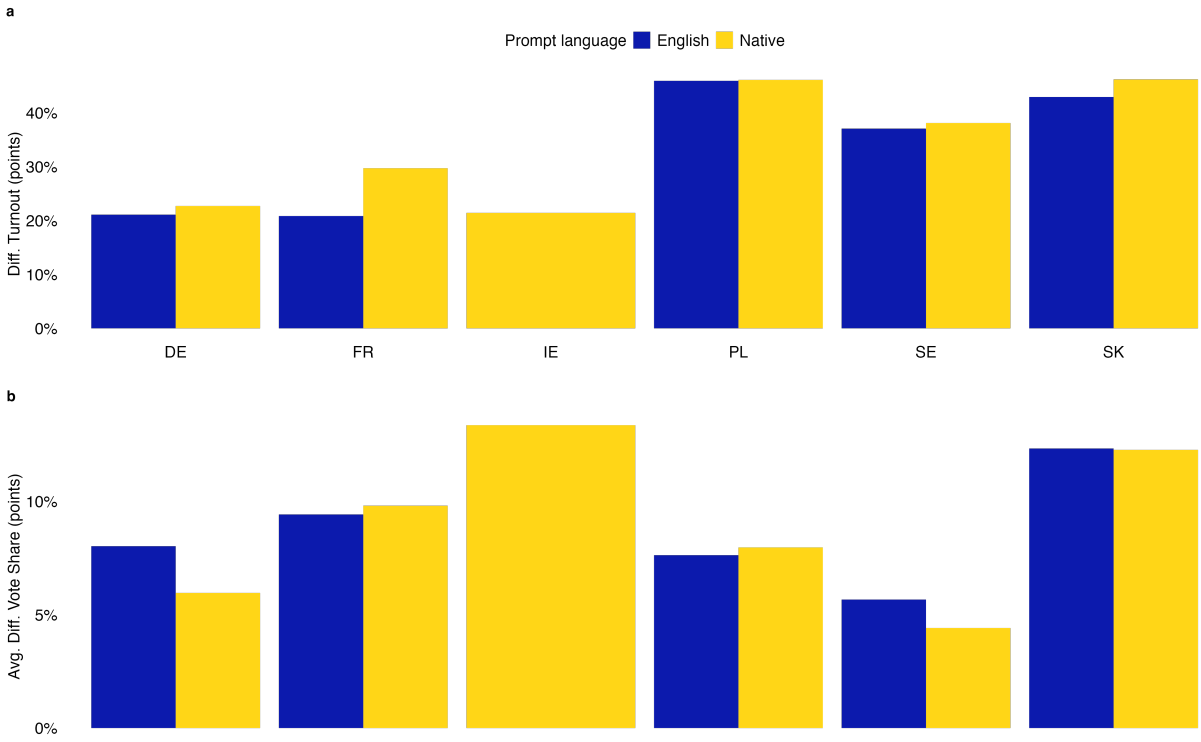


Figure 3.6: (Average) difference between actual turnout and party vote shares in the 2024 EU elections and GPT-4-Turbo's predictions (based on full prompt) **by prompt language**.

Differences in predictive performance depending on prompt content

When prompting GPT-4-Turbo in English with only demographic information about European citizens, the LLM tends to overestimate turnout even more (Figure 3.7a), and make even less accurate predictions of individual party vote shares than when prompted with additional attitudinal information for most countries (Figure 3.7b). Even in Belgium, where the full prompt led to an underestimation of turnout, GPT-4-Turbo overestimates turnout. For eight countries, predicted vote shares based on demographic information are closer to the actual result than those based on more detailed information. This includes Baltic states, some Eastern European countries, as well as Luxembourg and Malta. The apparent randomness of these results suggests an underlying randomness in when LLM-based predictions of voting behavior are correct, questioning the reliability of the method. Per-country-averages of absolute differences between predicted and actual vote shares for individual parties also have a lower variance when using only demographic information, suggesting that GPT-4-Turbo systematically misestimates vote shares regardless of the country or individual in question without additional information that would provide nuance (Figure A2.3).

Also when using native-language prompting, providing only demographic information leads

to vastly higher overestimations of turnout compared to the full set of information (Figure 3.8a) and larger differences to actual party vote shares (Figure 3.8b). While the difference between demographic and full prompt is especially large for German and French, the level of divergence from the actual result is generally higher for Polish and Slovak (for turnout), suggesting a systematic bias against those countries and languages. In other words, if provided with more, and attitudinal information about individuals, GPT-4-Turbo's predictions of voting behavior are better. GPT-4-Turbo is systematically worse at predicting voting behavior for Eastern European countries and/or countries with Slavic languages, regardless of prompt language or the amount of information provided in the prompt.

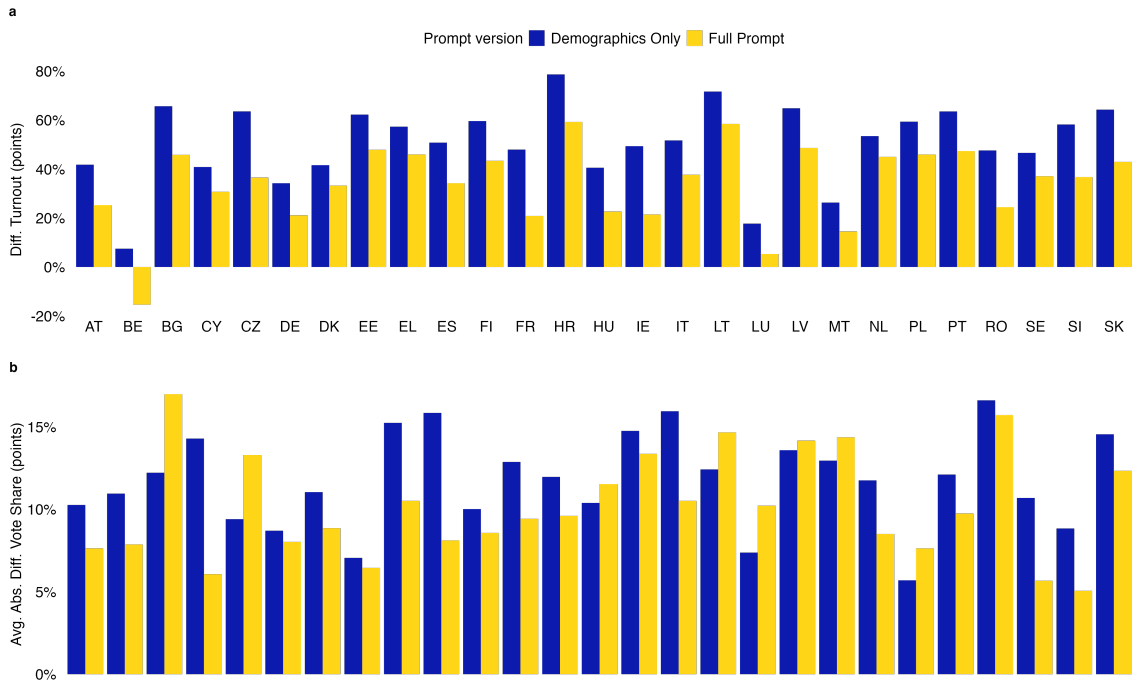


Figure 3.7: (Average) difference between actual turnout and party vote shares and predictions using GPT-4-Turbo (based on **English** prompt) **by prompt content**.

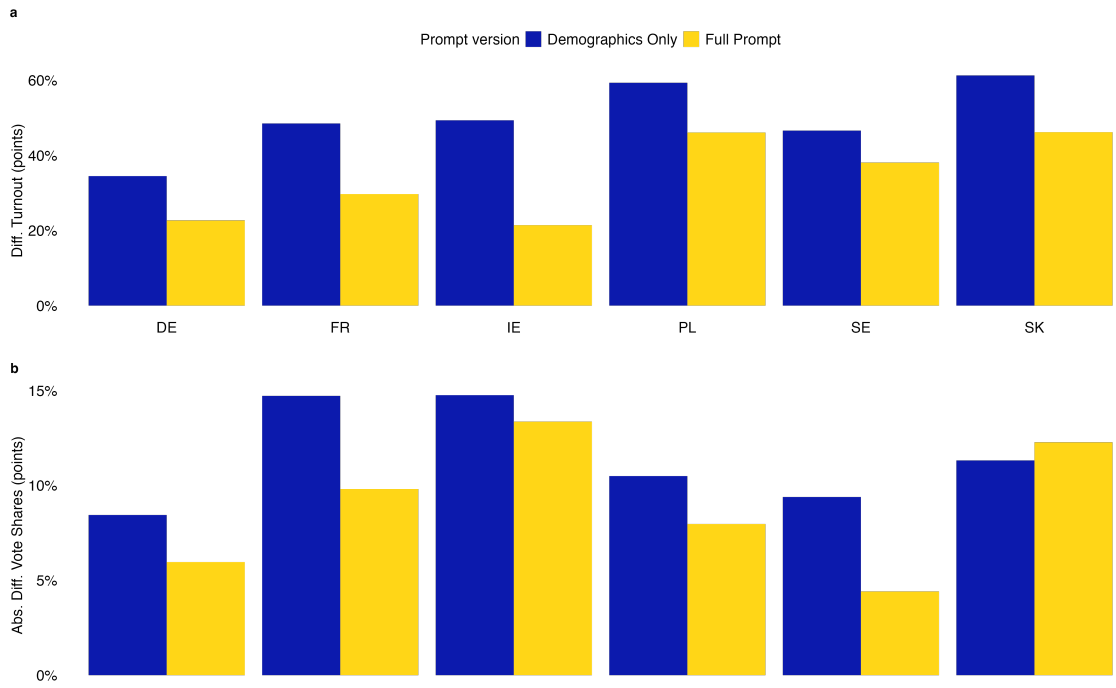


Figure 3.8: (Average) difference between actual turnout and party vote shares and predictions using GPT-4-Turbo (based on **native language** prompt) **by prompt content**.

Comparison to open-source models

Using the same prompts and model configuration, the open-source models Llama 3.1 and Mistral appear even less suitable for predicting public opinion based on synthetic samples. Compared to GPT, predictions using Llama 3.1 lead to larger overestimations and bigger contextual biases when it comes to turnout (see Appendix 2.6 for figures). However, Llama-based predictions are not as biased when it comes to vote shares (although similarly wrong on average). The same holds for predictions based on only demographic information, which are much worse for turnout, but not so much for vote shares. Llama exhibits even poorer predictive performance when prompted in native languages, which can be attributed to its more limited multilingual capacities.

Llama-based predictions overestimate turnout in every country and to an even larger extent than GPT, predicting an average of 95% (an overestimation of 46 percentage points). Echoing GPT, predictions are better for countries with high actual turnout and those countries with compulsory voting that are Western European using a dominant language. Predictive performance of party popularity is similarly poor as GPT in terms of winning party, party ranking, and average absolute differences to actual party vote shares. Biases against Eastern European countries and countries with Slavic or Baltic languages are more pronounced for predictions of turnout than party vote shares when using Llama than when using GPT. Despite the difference in patterns when it comes to the outcomes investigated (turnout vs. vote shares), these results suggest that biases against Eastern European countries are present regardless of the brand of LLM used. In contrast to GPT, prompting in a country's native language yields mixed results when it comes to predictions of turnout, and predictions of vote shares based on English-language prompting outperform those with native-language prompting for all countries, especially France and Poland. Overall, this

suggests that Llama's multilingual capacities are not as good as GPT's. While Llama has been trained on a broad range of languages, from our selection, only English, French, and German are officially supported use cases (Meta, n.d.). Thus, it is not surprising that its performance is weaker. Finally, providing only demographic information about individual voters leads to an even larger overestimation of turnout in all countries and higher divergences of predicted compared to actual party vote shares in most countries. While this pattern holds for native-language prompting regarding turnout predictions, the divergence to actual vote shares is not much different with or without attitudinal information.

Mistral largely did not follow the instruction of keeping the answer as short as possible, but instead either repeated the information contained in the prompt or stated that it was too difficult to make a definite prediction, both resulting in a disproportionate amount of completions lacking a vote choice, i.e., missing values (see Appendix 2.6 for details). Analyses based on the remaining data would neither be meaningful nor comparable to the other models. We conclude that Mistral cannot be used for generating synthetic samples for public opinion prediction in a similar manner as other models.

3.4 Discussion and Conclusion

Our results show that overall, LLMs fail at predicting turnout and party vote shares in the 2024 European elections based on synthetic samples of the voting population – they overestimate turnout and are largely unable to accurately predict the winner, rank ordering, or individual party vote shares. Only providing socio-demographic information about individual voters further worsens the results, casting severe doubts on the feasibility of using LLM-based synthetic samples as a supplement, let alone substitution, of detailed survey data. Finally, the LLMs are especially bad at predicting voting behavior for Eastern European countries and countries with native Slavic languages, regardless of language used or the amount of information provided in the prompt, suggesting systematic contextual biases.

Predicting political attitudes and behaviors in multi-party contexts is more complex than in the U.S. two-party system (von der Heyde et al., 2025), which most previous studies on LLM-synthetic samples investigated. As our findings show, predictions of future public opinion based on off-the-shelf LLMs do not live up to the hope of being a resource-efficient alternative in just any context, as they are not able to capture the complex mechanisms behind public opinion formation equally across contexts if these mechanisms are not featured in the training data (McCoy et al., 2023). Considering *what purpose* LLMs were trained to fullfill along with *how* they were trained to fulfill it (McCoy et al., 2023) and the training's *temporality* can help explain why LLMs fail in this task. Previous research has found that LLMs are better at retrodiction, i.e., retroactively imputing past opinions, than at predicting attitudes on new survey items, policy issues or events that occurred past its training data. When predicting, LLMs seem to instead generalize along broad ideological lines without regard for nuance (Kim & Lee, 2023; Sanders et al., 2023) – something that is reflected in the response distributions, which are different from human-generated survey data, often being less diverse (Bisbee et al., 2024; Dominguez-Olmedo et al., 2024; Hämäläinen et al., 2023). This should not be surprising, considering LLMs were trained to predict the most likely next words following a string of words. Put differently, they will output words that follow previous words with a high probability based on their training corpus. In instances where information about the task (here, predicting a specific, unobserved election outcome) occurs in the training data with low probability, LLMs will output words that, in its training corpus, are related

to cues in the input prompt with a high probability, even though the context of the input might be a completely different one (McCoy et al., 2023). Furthermore, traditional polls, whether regarded as input knowledge sources or output benchmarks for LLMs, provide a snapshot in time, both in terms of a population's structure and its attitudes. Public opinion, however, is volatile, and while voting behavior has certain stable long-term predictors (e.g., Rattinger & Wiegand, 2014), it is susceptible to shock events. Such short-term contextual changes and ensuing shifts in party popularity and strategic voting cannot be captured by LLMs with a knowledge cutoff far ahead of the event they are supposed to predict. For example, ahead of the 2024 European elections, several scandals within the far-right parliamentary groups dominated the news cycles and debates. In the specific case of elections, comparing the LLM-predictions' closeness to the results in the previous elections may shed light on how much the LLM's predictions are based on past patterns as opposed to new developments. Further, it may be worth exploring whether fine-tuning LLMs on recent, pertinent news and social media debates would improve results, as others have done for BERT with media diets (Chu et al., 2023) or Llama with Twitter data (Ahnert et al., 2025). Such content could simply be accessed and analyzed directly, but LLMs may still provide an advantage for aggregating and analyzing such digital trace data.

Our findings are consistent with our hypothesis on contextual biases in an LLM's data-generating process not just vis-à-vis the United States, but also within Europe: GPT-4-Turbo's and Llama's predictions more typically match the voting behavior of Western European countries, which likely can be explained by their larger linguistic and political presence in Europe and presence in the training data. These discrepancies in the digital divide are mirrored in our findings, which suggest that the LLMs are worse at predicting Eastern European voting behavior regardless of prompt language (and information provided in the prompt). The observed ambiguity of prompt language impact on predictive accuracy in our study mirrors existing findings, with some research suggesting prompting in a culture-specific language could mitigate biases to some extent (W. Wang et al., 2024), but other research finding consistent bias across languages (Durmus et al., 2024; Hartmann et al., 2023; Öztürk et al., 2025).

Considering the impact of information contained in the prompt on prediction quality, our findings suggest that demographic information alone is insufficient for accurately estimating complex individual-level attitudes. Our cross-national and cross-lingual comparison thus confirms previous case studies using various GPT models (Hwang et al., 2023; Lee et al., 2023; von der Heyde et al., 2025): There appear to be trade-offs between model sophistication and quality (Lee et al., 2023; Li et al., 2024), with newer models performing comparatively better given attitudinal information, but worse than older models given only demographic information. The fact that providing (general) attitudinal information in the prompt leads to better estimates of voting behavior gives rise to two considerations. It suggests that by adding even more (attitudinal) information about voters (in the European elections case, this might be, e.g., party identification, satisfaction with the national government, salience of and attitude towards issues such as immigration, economic growth, or climate change, and voting behavior in the last election), predictions might further improve. However, in our study, such data was not available with the most recent Eurobarometer sample, once again highlighting the tradeoff between recency and detail of human samples on which LLM-synthetic samples can be based. This leads to the second point: if detailed individual attitudinal information is required for an LLM to make accurate predictions of voting behavior or other items of public opinion, then LLM-based synthetic samples provide little advantage for computational social scientists, as they still need to resort to surveys to obtain such information.

Nevertheless, LLM-based predictions have potential for improvement. Future research could

benefit from a political science perspective, including engaging with learnings from polling and election forecasting. In this context, panel survey data might provide additional advantages for research on LLM-synthetic samples. For example, comparing LLM predictions of, e.g., voting behavior based on pre-election survey data with post-election survey data could give insights into whether LLMs could substitute post-election surveys. Experiments with different pre-event waves could also reveal where the "survey data cutoff" point is for LLMs to succeed in this task. However, survey data is not free from errors, potentially challenging the appropriateness of using it as a benchmark for LLMs, as opposed to observational data. While fine-grained, individual-level observational data (e.g., actual as opposed to reported voting behavior) is hardly available for most social science concepts, future research could evaluate whether survey or LLM output better mirrors aggregate real-world phenomena (e.g., election results), and which factors influence the difference of either prediction to such real-world observations.

Regarding the generalizability of our findings to different LLMs, the biases GPT-4-Turbo exhibits are mirrored in the open-source model Llama. These results suggest a systematic underlying issue in LLM training and fine-tuning that needs to be addressed (McCoy et al., 2023), and highlight the need for better multilingual and multicultural capacities. There are indications that certain models, such as ERNIE (an LLM trained on a balanced mix of English and non-English data) exhibit less cultural bias (W. Wang et al., 2024). Other research suggests that base models are less biased in terms of political orientation, at least on the aggregate level (Rozado, 2024), and less sensitive to bias-inducing prompting (Tjuatja et al., 2024) – however, at the cost of less coherence (Rozado, 2024). This may suggest that political biases in LLMs are created in the fine-tuning stages, not as a result of biased training data (Rozado, 2024). However, politically "neutral" fine-tuning may bring out biases created due to unbalanced training corpora, and even the active alignment against explicit biases might inadvertently exacerbate covert stereotypes (Hofmann et al., 2024; Li et al., 2024). Ultimately, "neutrality" is in the eye of the beholder, and alignment processes implicitly mirror the value systems of the people performing the alignment (Kirk et al., 2024). Transparency and diversity in the training and fine-tuning processes can guide the development of fairer and more accurate LLMs for computational social science applications (Huckle & Williams, 2025; McCoy et al., 2023). The increased development of LLMs for typically underrepresented languages, such as the TrustLLM (TrustLLM, n.d.) or No Language Left Behind (NLLB Team et al., 2024) projects, point in this direction. In addition, since model architecture and training data both influence LLM behavior (McCoy et al., 2023), future research should investigate how different LLMs' outcomes change when provided with different training corpora. However, only a few large companies have the resources to train LLMs, making such experiments largely inaccessible to the scientific community. The importance of an LLM's architecture, and with it, purpose, is evident in our results, where Mistral proved to be entirely unsuitable for the task at hand. This shows that researchers need to seriously consider the intended use cases of off-the-shelf LLMs (McCoy et al., 2023) and potentially customize models for their needs (e.g., Holtdirk et al., 2024). While LLMs may be *general*-purpose tools, that does not mean they are, by default, suitable for highly *specific* tasks such as public opinion prediction.

In conclusion, our findings emphasize the limited applicability of popular, state-of-the-art LLMs to public opinion prediction. The prediction of attitudes and behaviors relating to events that go beyond LLM training data is what would benefit most from LLMs' efficiency, but such predictions, which necessarily are based on past training and population data, largely fail. Moreover, LLM prediction accuracy is unequally distributed across countries and languages, even when using individual-level prompting information. Finally, improving LLM predictions requires detailed attitudinal information about individuals. Practitioners need to carefully examine the applicability

of LLM-synthetic sampling in the specific target context before drawing any conclusions, so as to not reproduce existing biases. For researchers, our findings point to the need to improve LLMs' training and fine-tuning to mitigate biases and inequalities against specific populations.

References

Aaru. (n.d.). Aaru: Rethinking the Science of Prediction. Retrieved November 21, 2024, from https://aaruaaru.com

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies [Place: Honolulu, Hawaii, USA]. *Proceedings of the 40th International Conference on Machine Learning*.

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2025). Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 15–36. https://doi.org/10.1609/icwsm.v19i1.35801

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 1–15. https://doi.org/10.1017/pan.2023.2

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023, September). Which Humans? https://doi.org/10.31234/osf.io/5b26t

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

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*, 1–16. https://doi.org/10.1017/pan.2024.5

Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for Market Research. https://doi.org/10.2139/ssrn.4395751

Braun, D., & Schäfer, C. (2022). Issues that mobilize Europe. The role of key policy issues for voter turnout in the 2019 European Parliament election. *European Union Politics*, *23*(1), 120–140. https://doi.org/10.1177/14651165211040337

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. *Proc. First Workshop Cross-Cult. Considerations NLP*, 53–67. https://doi.org/10.18653/v1/2023.c3nlp-1.7

Chu, E., Andreas, J., Ansolabehere, S., & Roy, D. (2023, March). Language Models Trained on Media Diets Can Predict Public Opinion [arXiv:2303.16779 [cs]]. https://doi.org/10.48550/arXiv.2303.16779

Chua, G. (2024). An AI polling startup makes its predictions for the 2024 US election. *Semafor*. Retrieved November 21, 2024, from https://www.semafor.com/article/11/04/2024/an-ai-polling-startup-polls-bots-predicts-harris-will-win

Delve AI. (n.d.). AI Persona Generator. Retrieved November 21, 2024, from https://www.delve.ai/ai-persona-generator

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2024). Questioning the Survey Responses of Large Language Models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information Processing Systems* (pp. 45850–45878, Vol. 37). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/515c62809e0a29729d7eec26e2916fc0-Paper-Conference.pdf

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A.,

Korenev, A., Hinsvark, A., Rao, A., Zhang, A., ... Zhao, Z. (2024, August). The Llama 3 Herd of Models [arXiv:2407.21783 [cs]]. https://doi.org/10.48550/arXiv.2407.21783

Durmus, E., Nguyen, K., Liao, T., 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. (2024). Towards Measuring the Representation of Subjective Global Opinions in Language Models. *First Conference on Language Modeling*. https://openreview.net/forum?id=zl16jLb91v

European Commission, B. (2024). Eurobarometer 99.4 (2023) [Published: GESIS, Cologne. ZA7997 Data file Version 1.0.0, https://doi.org/10.4232/1.14167]. https://doi.org/10.4232/1.14167

European Parliament. (n.d.). 2024 European election results [Last Modified: 2024-08-20T09:36:44.863810164Z Publisher: http://www.europarl.europa.eu/portal/en]. Retrieved August 21, 2024, from https://results.elections.europa.eu/en/index.html

EuroVoc. (n.d.). Browse by EuroVoc - EUR-Lex [Usr_lan: en]. Retrieved August 26, 2024, from https://eur-lex.europa.eu/browse/eurovoc.html?params=72%2C7206

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

Giebler, H., & Wagner, A. (2015). Contrasting First- and Second-Order Electoral Behaviour: Determinants of Individual Party Choice in European and German Federal Elections. *German Politics*, *24*(1), 46–66. https://doi.org/10.1080/09644008.2014.949684

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. *Proc. 2023 CHI Conf. Hum. Factors Comput. Syst.*, 1–19. https://doi.org/10.1145/3544548.3580688

Hartmann, J., Schwenzow, J., & Witte, M. (2023, January). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation [arXiv:2301.01768 [cs]]. Retrieved March 28, 2023, from http://arxiv.org/abs/2301.01768

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, *633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

Holtdirk, T., Assenmacher, D., Bleier, A., & Wagner, C. (2024, October). Fine-Tuning Large Language Models to Simulate German Voting Behaviour (Working Paper). https://doi.org/10.31219/osf.io/udz28

Huckle, J., & Williams, S. (2025). Easy Problems that LLMs Get Wrong. In K. Arai (Ed.), *Advances in Information and Communication* (pp. 313–332). Springer Nature Switzerland.

Hwang, E., Majumder, B., & Tandon, N. (2023, December). Aligning Language Models to User Opinions. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Find. Assoc. Comput. Linguist. EMNLP 2023* (pp. 5906–5919). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.393

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023, October). Mistral 7B [arXiv:2310.06825 [cs]]. https://doi.org/10.48550/arXiv.2310.06825

Jiang, S., Wei, L., & Zhang, C. (2024, November). Donald Trumps in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models

[arXiv:2411.01582 [econ]]. Retrieved November 19, 2024, from http://arxiv.org/abs/2411.01582

Kim, J., & Lee, B. (2023, November). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction [arXiv:2305.09620 [cs]]. Retrieved January 23, 2024, from http://arxiv.org/abs/2305.09620

Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, *6*(4), 383–392. https://doi.org/10.1038/s42256-024-00820-y

Kleinberg, B. (2024, May). Rgpt3: Making requests from R to the GPT API. https://doi.org/10.5281/zenodo.7327667

Lee, N., An, N., & Thorne, J. (2023). Can Large Language Models Capture Dissenting Human Voices? *Proc. 2023 Conf. Empir. Methods Nat. Lang. Process.*, 4569–4585. https://doi.org/10.18653/v1/2023.emnlp-main.278

Levanti, M., & Verret, C. (2024, September). The rise of synthetic respondents in market research: Why some will make it and some will fake it. Retrieved November 21, 2024, from https://nielseniq.com/global/en/insights/education/2024/the-rise-of-synthetic-respondents/

Li, B., Haider, S., & Callison-Burch, C. (2024). This Land is Your, My Land: Evaluating Geopolitical Bias in Language Models through Territorial Disputes. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3855–3871. https://doi.org/10.18653/v1/2024.naacl-long.213

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, *9*(8). https://doi.org/10.18637/jss.v009.i08

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023, September). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve [arXiv:2309.13638 [cs]]. Retrieved September 18, 2024, from http://arxiv.org/abs/2309.13638

Mendoza, D. (2024). AI polling company defends wrong predictions on the US election. *Semafor*. Retrieved November 21, 2024, from https://www.semafor.com/article/11/06/2024/ai-startup-aaru-defends-using-artificial-intelligence-for-polling

Meta. (n.d.). Meta-Llama-3.1-8B-Instruct. Retrieved September 16, 2024, from https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

MistralAI. (2023, September). Mistral 7B [Section: news]. Retrieved September 17, 2024, from https://mistral.ai/news/announcing-mistral-7b/

MistralAI. (n.d.). Mistral-7B-Instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

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

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., … Wang, J. (2024). Scaling neural machine translation to 200 languages. *Nature*, *630*(8018), 841–846. https://doi.org/10.1038/s41586-024-07335-x

OpenAI. (n.d.). Models – GPT-4 Turbo and GPT-4. Retrieved March 26, 2024, from https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

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 Number: 6]. https://doi.org/10.48550/ARXIV.2303.08774

Öztürk, I. T., Nedelchev, R., Heumann, C., Arias, E. G., Roger, M., Bischl, B., & Aßenmacher, M. (2025). How Different is Stereotypical Bias Across Languages? In R. Meo & F. Silvestri (Eds.), *Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (pp. 209–229). Springer Nature Switzerland.

Pezeshkpour, P., & Hruschka, E. (2024, June). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006–2017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.130

Qi, W., Lyu, H., & Luo, J. (2024, July). Representation Bias in Political Sample Simulations with Large Language Models [arXiv:2407.11409 [cs] version: 1]. https://doi.org/10.48550/arXiv.2407.11409

Qu, Y., & Wang, J. (2024). Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, *11*(1), 1095. https://doi.org/10.1057/s41599-024-03609-x

R Core Team. (2024, July). R: The R Project for Statistical Computing. Retrieved July 2, 2024, from https://www.r-project.org/

Rattinger, H., & Wiegand, E. (2014, May). Volatility on the Rise? Attitudinal Stability, Attitudinal Change, and Voter Volatility [_eprint: https://academic.oup.com/book/0/chapter/151985894/chapter-ag-pdf/44973716/book_7270_section_151985894.ag.pdf]. In *Voters on the Move or on the Run?* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199662630.003.0013

Rozado, D. (2024). The political preferences of LLMs (T. Zhang, Ed.). *PLOS ONE*, *19*(7), e0306621. https://doi.org/10.1371/journal.pone.0306621

Sabbati, G., & Grosek, K. (2023). 2024 European elections: National rules. https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2023)754620

Sanders, N. E., Ulinich, A., & Schneier, B. (2023). Demonstrations of the Potential of AI-based Political Issue Polling [Publisher: The MIT Press]. *Harvard Data Science Review*, *5*(4).

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023, July). Whose Opinions Do Language Models Reflect? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 29971–30004, Vol. 202). PMLR. https://proceedings.mlr.press/v202/santurkar23a.html

Synthetic Users. (n.d.). Synthetic Users. Retrieved November 21, 2024, from https://www.syntheticusers.com

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., & Neubig, G. (2024). Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design [_eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00685/2468689/tacl_a_00685.pdf]. *Transactions of the Association for Computational Linguistics*, *12*, 1011–1026. https://doi.org/10.1162/tacl_a_00685

TrustLLM. (n.d.). TrustLLM: Democratizing Trustworthy and Factual Large Language Model Technology for Europe. Retrieved August 21, 2024, from https://trustllm.eu/

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R [Number: 3]. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

von der Heyde, L., Haensch, A.-C., & Wenz, A. (2025). Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice [_eprint: https://doi.org/10.1177/08944393251337014]. *Social Science Computer Review*, *0*(0), 1–23. https://doi.org/10.1177/08944393251337014

W3Techs. (2024, July). Usage Statistics and Market Share of Content Languages for Websites, July 2024. Retrieved July 2, 2024, from https://w3techs.com/technologies/overview/content_language

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-025-00986-z

Wang, P., Zou, H., Yan, Z., Guo, F., Sun, T., Xiao, Z., & Zhang, B. (2024, September). Not Yet: Large Language Models Cannot Replace Human Respondents for Psychometric Research. https://doi.org/10.31219/osf.io/rwy9b

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-t., Tu, Z., & Lyu, M. (2024, August). Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6349–6384). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.345

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 [Number: 43]. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wikipedia. (2024, August). Languages of Europe [Page Version ID: 1239449282]. Retrieved August 21, 2024, from https://en.wikipedia.org/w/index.php?title=Languages_of_Europe&oldid=1239449282

4 AIn't Nothing But a Survey? Using Large Language Models for Coding German Open-Ended Survey Responses on Survey Motivation

4.1 Introduction

The recent development and wider accessibility of large language models (LLMs) have spurred discussions about how these language models can be used in survey research. Potential applications span the entire survey lifecycle, including using LLMs for questionnaire design and pretesting (e.g., Götz et al., 2023), conducting interviews (e.g., Cuevas et al., 2023), synthesizing or imputing respondent data (e.g., Argyle et al., 2023; Kim & Lee, 2023), or detecting non-human respondents in online surveys (e.g., Lebrun et al., 2024). Due to their linguistic capacities, including their adaptability to different topics, the detection of nuance, implicitness, and intent in low-information multilingual textual input, and flexibility in generating textual output, LLMs also offer promising potential for classifying open-ended survey responses, which often are short and do not provide explicit context. For example, using LLMs for coding free-text social media data has successfully been applied for efficiently capturing detailed public opinion data (Ahnert et al., 2025; Cerina & Duch, 2023) – an application that could be transferred to open-ended responses. Other popular semi-automated classification approaches for open-ended responses, such as support vector machines or random forests (e.g., Haensch et al., 2022; Landesvatter, 2024), are less adaptable across different languages and often require substantial expertise, pre-processing, and training data coded by humans (Landesvatter, 2024). Since LLMs could potentially eliminate the need for these time- and expertise-intensive requirements, it is possible that they are an efficient alternative for classifying open-ended responses in survey research. While researchers have begun to explore this application of LLMs (Landesvatter, 2024; Mellon et al., 2024; Rytting et al., 2023) and were largely successful, most of these studies have focused on English-language responses, responses relating to non-complex topics, or on single LLMs and prompting strategies. It is thus unclear to what extent existing findings generalize to other LLMs, prompting strategies, languages, and more complex topics. Furthermore, research has raised concerns about the reproducibility of LLM-generated output due to their non-deterministic design (Barrie et al., 2024), an issue that can extend to the coding of open-ended responses when it comes to the reliability of the coding, for example when new survey data is available. Overall, the exact conditions of the applicability of LLMs for coding open-ended survey data and the quality of these classifications, also compared to more established methods, have yet to be understood.

In this project, we are the first to investigate to what extent different LLMs can be used to code non-English (German) open-ended responses on survey motivation given a predefined set of

categories. We examine performance and reliability, and the dependency of these indicators on two factors – model selection and prompting approach. Specifically, we ask:

RQ1: Are there differences between LLMs regarding the performance and reliability of the coding?
RQ2: Are there differences between prompting approaches regarding the performance and reliability of LLM-based coding?

RQ2a: Does providing detailed descriptions of categories improve the performance and reliability of the coding?

RQ2b: To what extent does few-shot prompting impact the performance and reliability of the coding compared to zero-shot prompting?

RQ2b: Does fine-tuning an LLM on a subset of pre-coded response data improve the performance and reliability of the coding?

To do so, we contrast proprietary and open-source LLMs – GPT-4o, Llama 3.2, and Mistral NeMo, which are the most capable multilingual models of their respective families to date. We compare their category assignments when using zero-shot prompting (i.e., not providing examples) with and without category descriptions and few-shot prompting (i.e., providing exemplary classifications), and fine-tuning (i.e., further training of the LLM), and evaluate them against the codings of human experts. We also discuss the LLMs' performance in contrast to other classification methods reported in previous studies. By comparing the use of different LLMs and prompting approaches for classifying open-ended survey responses in German, our study uniquely contributes to the growing body of research about the conditions under which LLMs can be efficiently, accurately, and reliably leveraged in survey research and about the impact of LLM use on data quality.

4.2 Background

There are three main types of approaches to coding open-ended survey responses: traditional human coding, supervised machine learning methods, and the still-emerging use of LLMs, each with distinct strengths and challenges. In this section, we review these methods, highlighting the potential of LLMs that yet needs to be explored.

In manual coding, human coders assign responses to predefined categories. While considered mostly accurate, this approach is time-intensive and costly, especially for large survey datasets or such with multiple open-ended questions (Haensch et al., 2022; Landesvatter, 2024). Costs are compounded when wanting to increase validity and reliability by having responses classified by several coders. These factors contribute to the sparseness of open-ended questions in survey instruments, despite such items allowing for deeper, authentic insights into how individuals think and act (Haensch et al., 2022).

Supervised methods attempt to address this resource-intensiveness by combining manual coding of a training dataset with machine learning algorithms, such as support vector machines (SVMs; Joachims, 2001) or gradient boosting (Schonlau & Couper, 2016). Applications to political (Grimmer & Stewart, 2013) and economic texts (Gentzkow et al., 2019) as well as other survey responses (Haensch et al., 2022; Schierholz & Schonlau, 2021) demonstrated their utility. But while these sophisticated approaches can somewhat reduce costs and time, they still require a substantial amount of human-coded data and expertise and computational resources for model

training in order to achieve satisfactory results, making them inefficient. They also struggle with short open-ended survey responses, which often lack sufficient context. In addition, they are usually only trained for one specific language and topic, making them not easily transferable across studies and less feasible for multilingual studies.

Transformer-based models, such as BERT, are able to capture nuanced relationships in text due to their ability to generate contextual embeddings. This offers improved classification performance for open-ended survey questions (Gweon & Schonlau, 2024; Meidinger & Aßenmacher, 2021). For example, Schonlau et al. (2023) demonstrated BERT's effectiveness for coding German-language survey questions, such as the GLES "most important problem" question. However, applying BERT to survey data poses similar challenges as supervised methods, as open-ended responses are often too short to utilize the models' full potential, and fine-tuning them to the specific types of (con)text requires expertise and computational resources (e.g., Schonlau et al., 2023). In addition, Schonlau et al. (2023) required the hand-coding of 80% of the data for training and validation, i.e., over 14,000 responses. Although effective, using BERT may thus not be very efficient.

While BERT is an *analytical* language model designed primarily for specific tasks like classification or entity recognition at the sentence or document level, modern-day *generative* large language models such as GPT-4 are designed to perform a broader range of generative and context-adaptive language processing tasks, including handling complex dialogs, summarization, and multilingual text generation. Such general-purpose LLMs thus show potential to address limitations of earlier approaches when applied to open-ended survey responses, like handling short responses when given only general information on their context, not necessarily requiring pre-coded data for training or fine-tuning, and being flexibly usable across languages. In addition, since off-the-shelf LLMs do not require large programming expertise, are relatively cost-effective, and can follow natural language instructions, they are more accessible to a broader group of survey researchers than other semi-automated methods. LLMs have brought promising advancements to labeling other types of social science text data, such as social media data and political texts, with studies finding that LLMs were at least on par or even outperformed supervised methods (Ahnert et al., 2025; Ornstein et al., 2024; Törnberg, 2024), making them applicable for substantive downstream analyses, like predicting public opinion (Ahnert et al., 2025; Cerina & Duch, 2023; Heseltine & Clemm von Hohenberg, 2024). Research specifically evaluating the applicability of LLMs for coding open-ended survey responses, however, continues to be scarce. In addition, LLMs' rapid evolution requires constant reevaluation of their precision and domain-specific applicability (Pangakis et al., 2023).

Rytting et al. (2023) tasked GPT-3 to code 7,500 English open-ended responses on keyword descriptions of U.S. partisans into binary and ternary categories. The LLM-based coding matched the (poor) performance of human crowdworkers and experts in terms of inter-coder agreement. It also came close to the performance of a supervised approach while needing substantially fewer labelled examples. Mellon et al. (2024) come to similar conclusions when testing a larger and more recent variety of open- and closed-source LLMs for coding several thousand open-ended responses to the "most important issue" question in the British Election Study into 50 categories. Benchmarked against a trained human coder, LLMs' accuracy of classifications varied between and within model families. Compared to a range of supervised approaches, the general-purpose LLMs performed much better, with BERT-based methods still outperforming SVMs.

Using LLMs for coding open-ended survey responses thus appears like a promising method for survey researchers. However, these studies represent a best-case scenario of relatively easy tasks, as they cover English-language data about standard societal and political issues that are

likely much-discussed in LLM training data and do not require much expertise for coding. Research on logical reasoning tasks suggests that LLMs tend to struggle with tasks that are comparably complex, but less commonly appearing in their training and alignment processes (McCoy et al., 2023). In addition, there is ample evidence that LLMs are biased against non-English language contexts in a variety of other tasks (e.g., Durmus et al., 2024; Johnson et al., 2022; Li et al., 2024; Wang et al., 2024). For example, Törnberg (2024) found that GPT-4 can be used for labeling non-English social media data, but Heseltine and Clemm von Hohenberg (2024) observed decreased speed and accuracy compared to English-language texts. Once again, these studies examined comparatively simple tasks, namely binary labeling of sentiment and political affiliation.

Beyond these limitations, there is competing evidence regarding specific LLM performance and prompting strategies: It is unclear whether all (families of) LLMs are equally suited for classifying open-ended responses. For example, most studies on using LLMs for coding social science text data investigated models of the GPT family, but came to conflicting conclusions regarding different model versions (e.g., Bosley et al., 2023 vs. Rytting et al., 2023 for GPT-3; Ornstein et al., 2024 vs. Heseltine and Clemm von Hohenberg, 2024; Törnberg, 2024 for GPT-4; Mellon et al., 2024 vs. Ahnert et al., 2025 for Llama). Considering proprietary vs. open-source model families, Mellon et al. (2024) found that the closed-source Claude models matched human coding best, followed by GPT-4, whereas Llama and PaLM performed much worse, and some other open-source LLM families were unable to complete the task at all.

Furthermore, existing research uses competing prompt designs. Some studies suggest zero-shot prompting (i.e., not providing examples for the labeling task, only the possible labels) is sufficient for labeling other types of short social science text data (Cerina & Duch, 2023), even in non-English languages (Törnberg, 2024). In contrast, studies applying LLMs specifically to open-ended survey responses used few-shot prompting (Mellon et al., 2024; Rytting et al., 2023). In this approach, the authors included the coding scheme along with three examples in the prompt, sometimes supplemented by detailed category descriptions. Halterman and Keith (2024) found that including more detailed definitions of the categories and positive as well as negative examples had a positive impact on labeling quality. However, Mellon et al. (2024) report that providing a full coding guide appeared to "distract" the LLMs. Finally, Mellon et al. (2024) suggest that fine-tuning, i.e., re-training LLMs on pre-labeled survey responses would likely further improve results. Ahnert et al. (2025) successfully used fine-tuning, albeit not for open-ended survey data.

Given this scarce and competing evidence, it remains unclear whether and which existing findings about the applicability of LLMs for coding open-ended survey responses generalize. In this study, we seek to close this gap by testing different LLMs and prompting strategies for multi-class, single-label classification of a more specific topic in German open-ended survey data.

4.3 Data and Methods

Open-ended survey data and coding scheme

In order to test the applicability of LLMs for coding German-language open-ended survey responses, we use data from a German probability-based mixed-mode panel, the GESIS Panel.pop Population Sample (Bosnjak et al., 2018; GESIS, 2024). Randomly sampled from municipal population registers, the panel includes over 5,000 respondents and covers the population of German-speaking permanent residents of Germany aged 18+. Participants are invited to the 20-minute survey waves bimonthly, receiving a prepaid incentive of five euros with every invitation. For

the years 2014 to 2020, the survey includes an annual, non-mandatory open-ended question on survey motivation. There, the panelists are asked to give their most, second most, and third most important reason for participating in the panel on three separate lines (see Appendix 3.2 for question wording). This questionnaire design leads to unidimensional answers usually containing only one category, making the item very favorable for coding (Haensch et al., 2022). Thus, while the response format should present an easy test case for LLMs, the specificity and complexity of the topic in terms of categorical dimensions, as well as the German language, present a harder task. The dataset contains a total of approximately 25,000 responses to the question on survey motivation across survey waves. For our study, we rely on a random sample of 20% of that data (5,072 responses) coded independently by two survey researchers (Cohen's kappa = 0.91, with remaining disagreements resolved by a more senior expert) based on a coding scheme for survey motivation adapted to the GESIS Panel.pop by the survey researchers (see Haensch et al., 2022, for details). The human codes are not necessarily required for employing LLMs (see below for a discussion of prompting approaches), but serve as a ground truth to compare the LLM-based classifications to. Indeed, when not fine-tuning an LLM, using it would require only a fraction of the human-coded examples necessary for training traditional supervised methods – for example, Haensch et al. (2022) used 5,000 human-coded responses to train an SVM.

For the LLM-based classifications, we use the same coding scheme as was used by the human coders. It spans 22 categories, featuring both intrinsic, extrinsic, and survey-related reasons for motivation (Haensch et al., 2022; Porst & von Briel, 1995). It also includes catch-all categories: No reason captures explicit statements of not having a reason for participation, "don't know"s, as well as non-meaningful fillers such as "???". In contrast, Other contains meaningful statements that cannot be assigned to any other category. For English translations of the categories, see Figure 4.1. A more detailed coding scheme with definitions and examples for all categories and their groups can be found in Appendix 3.1.

LLM selection and configuration

We test and compare powerful and popular LLMs of three different model families that are state-of-the-art at the time of writing. Models of one of the industry leaders, OpenAI, are popularly used by the public and researchers without large computational expertise due to their user-friendly accessibility. Despite OpenAI's lack of transparency and reproducibility as a proprietary provider (Palmer et al., 2023), it thus is reasonable to include one of their models in our research as a realistic use case. GPT-4o (GPT henceforth) is OpenAI's flagship model at the time of writing, which, according to the developers, features considerable improvements in non-English languages over earlier versions, while being more time- and cost-efficient (OpenAI, 2024a, 2024b). It is also supposed to be more capable of domain-specific or complex tasks and detailed labeling.

In line with calls for accessible and reproducible AI research (e.g., Spirling, 2023; Weber & Reichardt, 2023), we also test two open-source LLMs. These are downloaded and run locally, ensuring sensitive data remains private and is not shared with third parties. This is crucial as open-ended responses may inadvertently contain personal information, such as addresses, risking re-identification.[1] Running LLMs locally also ensures reproducibility by using stable model versions, unaffected by updates to cloud-based APIs (Spirling, 2023). Llama-3.2-3B-Instruct

[1] To ensure a similar level of privacy for the proprietary GPT model, we (fine-tune and) run it on AzureOpenAI, which provides private instances of GPT models on European servers.

(Llama henceforth) is the more capable of the two multilingual LLMs of Meta's Llama 3.2 suite, the most recent and powerful one at the time of writing (Meta, 2024a, 2024b). While the open-source suite also features larger models (11B and 90B), those are not optimized for multilingual dialog and not available in Europe, making them infeasible for the project at hand and international survey research more broadly. Mistral-NeMo-Instruct-2407 (Mistral henceforth) is the most recent multilingual model by the European open-source developers Mistral. It is specifically designed for global, multilingual applications (MistralAI, 2024a, 2024b) and supposed to be particularly strong in, among other languages, German. We access these models via the Huggingface platform (Meta, 2024b; MistralAI, 2024a).

To investigate the exact conditions under which LLMs can be used to code German open-ended survey responses, we employ different approaches.

Zero-shot prompting: In the least supervised approach, we simply ask the LLMs to classify the open-ended responses without any additional information apart from the coding scheme (i.e., no examples or definitions of responses belonging to the specific categories).

Zero-shot prompting with category descriptions: Along with the coding scheme, we provide the LLMs with definitions for each category.

Few-shot prompting: In few-shot prompting, an LLM is given a few examples to guide its output along with the coding scheme, providing an efficient alternative to training the LLM with task-specific data. To test how few-shot prompting impacts the performance of LLMs for open-ended response classification, we provide the LLMs with one example response per category (so 22 examples in total) in the prompt. The examples are randomly selected from the examples featured in the coding scheme, containing actual answers featured in the dataset of responses to be classified. They are presented in random order in the prompt. The examples are not removed from the classification dataset.

Fine-tuning: Fine-tuning involves further training the model on a smaller, domain-specific dataset to improve its performance on particular tasks. While less efficient because of the need for more human-coded training examples, fine-tuned LLMs might yield more accurate results than using LLMs out-of-the-box. Exploring whether fine-tuning a model on humanly pre-coded response data thus helps understand LLMs' potential in classifying open-ended responses.

However, depending on the LLM, fine-tuning requires even more extensive computing resources. This is not only a limitation for practitioners, but also for our test case. We therefore select only GPT-4o for fine-tuning, due to its straightforward and easily available fine-tuning services, making it a likely choice for researchers wishing to employ this approach. We fine-tune the LLM by splitting the dataset into a training and a test subset. As is common for fine-tuning tasks, we randomly select 80% of responses of each category based on the human classification (4,048 in total)[2] for training the LLM before asking it to classify the remaining 1,024 responses using the zero-shot prompt. Results for the fine-tuned approach thus reflect the LLMs' performance on the test set alone. We specify four epochs[3] for fine-tuning, i.e., four iterations through the training data, and use default values for batch size (the number of examples used in a single training pass; around 0.2% of the training dataset, ten in our case) and learning rate (rate at which the LLM

[2]We train the LLM with the zero-shot prompt including the responses in their raw form as input, not correcting any spelling mistakes or similar. As output, we use the desired completion format (see prompt design).

[3]Using four epochs for fine-tuning is a deliberate choice, balancing generalization and task-specific adaptation. While this is not the default, it represents a compromise between the lower range typically sufficient when using validation sets (i.e., 1–2 epochs) and OpenAI's recommendation to increase the number of epochs for tasks with a small set of ideal outputs, such as classification.

updates its weights (i.e., internal settings) based on the new data, balancing between learning too slowly, risking inefficiency, and too quickly, risking instability). Appendix 3.7 reports the loss and token accuracy curves of the fine-tuning process.

Since we want to maximize reliability and the task of coding responses according to a set of predefined categories does not require creativity but consistency, we set the LLM temperature to 0, thereby flattening the LLM's underlying probability function to produce more deterministic outputs. For best comparability, we use the same temperature for all models, leaving all other parameters at model default.

Prompt design

We tell LLMs to impersonate a survey expert classifying open-ended responses and instruct them to assign each response to exactly one category. The order of categories (and their descriptions in the detailed approach, and examples in the few-shot approach, respectively) is randomized in each prompt to avoid any biases due to order effects (Brand et al., 2023; Pezeshkpour & Hruschka, 2024). To minimize missing values, we ask the LLMs to make a best guess in difficult cases. We instruct the LLMs to report the response along with its classification. Finally, to avoid unnecessarily long answers, we ask the LLMs not to justify their response (as especially Mistral has been found to do previously, see, e.g., von der Heyde et al., 2024)[4], but do not specify a maximum output length. Figure 4.1 shows an English translation of the prompt. In line with the language of the responses they are being asked to classify, we prompt the LLMs in German, including the instructions and coding scheme. The original German version of the prompt, as used in the study, can be found in Appendix 3.3.

We prompt each survey response separately and with refreshed LLM memory, to ensure that responses are classified independently of one another. We therefore specify the task directly in the main prompt (not the system prompt), thereby repeating the task for every open-ended response to be classified. Before we feed the full dataset to the LLMs, we test each LLM with only 15 responses to determine its general capacity to fulfill the task. We run each query twice per LLM to be able to evaluate its reliability. All data is generated in November 2024, except the classifications obtained from the fine-tuned version of GPT, which is generated in January 2025.

Analysis

We extract each LLMs' classifications of the open-ended responses and analyze their performance and the resulting descriptive distributions. Benchmarking against the human-generated classifications, we analyze the LLMs' classification performance overall and per category. Because our case is one of multiclass-classification and the benchmark categories are unevenly distributed (see Figure 4.4, we use macro F1 scores[5] as our primary overall performance metric (Hand et al., 2024). In imbalanced datasets, regular F1 scores can be misleading if an LLM tends to assign the modal category. Macro F1 addresses this by averaging across the per-category F1 scores, giving equal weight to minority categories.

If an LLM failed to classify a response to exactly one category (i.e., it did not assign a

[4] Chapter 3 of this dissertation.

[5] Generally, F1 scores range from 0 to 1, with higher values indicating better predictive performance. For a more detailed description, see Appendix 3.5.

You are a survey expert classifying open-ended responses to the question why individuals participate in a survey. Assign these reasons for participating to exactly one of the following categories.

The categories are:
INTEREST: [Description]
CURIOSITY: [Description]
LEARNING: [Description]
TELL OPINION: [Description]
INFLUENCE: [Description]
INCENTIVE: [Description]
FUN: [Description]
ROUTINE: [Description]
DUTIFULNESS: [Description]
HELP SCIENCE: [Description]
HELP POLITICIANS: [Description]
HELP SOCIETY: [Description]
HELP, NOT FURTHER SPECIFIED: [Description]
BREVITY: [Description]
ANONYMITY: [Description]
PROFESSIONALISM: [Description]
RECRUITMENT: [Description]
RECRUITER: [Description]
OTHER SURVEY CHARACTERISTICS: [Description]
IMPORTANCE IN GENERAL: [Description]
OTHER: [Description]
NO REASON: [Description]

Make your best guess, even if it is hard.
Respond in the following format: Reason for participating | CATEGORY.
Do not give an explanation for your classification, but return only the reason for participating and your classification.

Examples:
[Example reason | CATEGORY 1]
[Example reason | CATEGORY 2]
[...]
[Example reason | CATEGORY 22]

Classify the following reason for participating:

[open-ended response]

Figure 4.1: English translation of prompt used for LLM-based classifications of the open-ended survey question.
Categories and, in the detailed approach, descriptions (green font) were randomized across individual queries. In the few-shot approach, examples (blue font) were randomly selected, the selection being held constant, but presented in random order across queries. For details of descriptions and examples used, see Appendix 3.1.

category or assigned more than one category), the output is recorded as missing (i.e., an explicit category called "NA") but retained for the analysis. This approach avoids artificially inflating the F1 scores for categories where most responses were not classified, but the remainder classified correctly,[6] and allows us to investigate the reliability of missing classifications. To facilitate comparison to other studies and classification methods, we report additional metrics (weighted F1, accuracy, intraclass correlation coefficients, Cohen's kappa) in Appendix 3.5. Since Haensch et al. (2022) previously tested an SVM on the same data, we are able to compare LLM performance to that of a supervised approach without explicitly having to employ that approach ourselves (see the Discussion section). To do so, we calculate the median F1 score as the unweighted median across categories. We then compare the distribution of coding scheme categories across LLMs and prompting approaches and to the distribution of the human-coded benchmark data. We also report the frequency and categorical distribution of the responses each LLM fails to classify as well as the reason for failure (see Appendix 3.4). For all analyses, we rely on the first iteration of classifications per LLM and prompting approach, independent of whether this iteration exhibited better or worse performance than the second one, in order not to bias our results by selecting on performance.

To assess the LLMs' reliability, we calculate the ICC for two-way agreement between the two iterations of classifications per LLM and prompting approach.

Data (pre-)processing, classification (for GPT), and analyses are conducted in R (version 4.3.2, R Core Team, 2024), especially using the packages *AzureAuth* (Ooi et al., 2019), *caret* (Kuhn, 2008), *irr* (Gamer et al., 2019), and *tidyverse* (Wickham et al., 2019). Classifications from Llama and Mistral are obtained using Python, especially using the packages *accelerate* ("Accelerate", n.d.), *huggingface_hub* ("Hub client library", n.d.), *pandas* (McKinney, 2010), *PyTorch* (Paszke et al., 2019), *tqdm* (da Costa-Luis et al., 2024), and *transformers* (Wolf et al., 2020).

4.4 Results

Differences between LLMs

Performance. We first compare differences between LLMs in classification performance overall (macro F1) and per category (F1). Across prompting approaches, classification performance is much better when using GPT than when using Mistral, which still has a slight edge over Llama (see Figure 4.2). GPT performance also fluctuates much less between prompting approaches (macro F1 around 0.7 for the three approaches that were examined for all three LLMs). Nevertheless, even using the best-performing prompting approach for an open-source LLM does not come near the GPT performance. Similar patterns emerge when considering other performance metrics (see Appendix 3.5).

All LLMs examined exhibit approximately the same performance patterns across categories (see Figure 4.3). They perform exceptionally well on the categories *incentive*, *interest*, and *fun* (macro F1 around 0.9), as well as on *anonymity*, *routine*, and *tell opinion*, and exceptionally poor (macro F1 between 0.02 and 0.3) on *no reason*, *non-identifiable/other*, and *other survey characteristics*. The LLMs thus perform very well on the three categories most commonly defined by the human coders, but not on the next two most common categories, which are non-substantive catch-all categories. For the remaining categories, performance tends to decrease along with

[6] The overall macro F1 scores exclude "missing" as an assigned category, as it is not meaningful or valid.

frequency of occurrence. The overall pattern is mirrored across types of reasons (extrinsic, intrinsic, survey-related): GPT's performance tends to be better than that of Llama and Mistral, which improves with few-shot prompting. There are some cases that stand out, which help explain the overall performative edge GPT has over the open-source models. GPT outperforms Llama and Mistral especially in *tell opinion*, *routine*, *importance in general*, *influence*, *dutifulness*, *curiosity*, and *professionalism*, and to a lesser extent also in the *help* categories, although Llama and especially Mistral improve under few-shot prompting.[7]

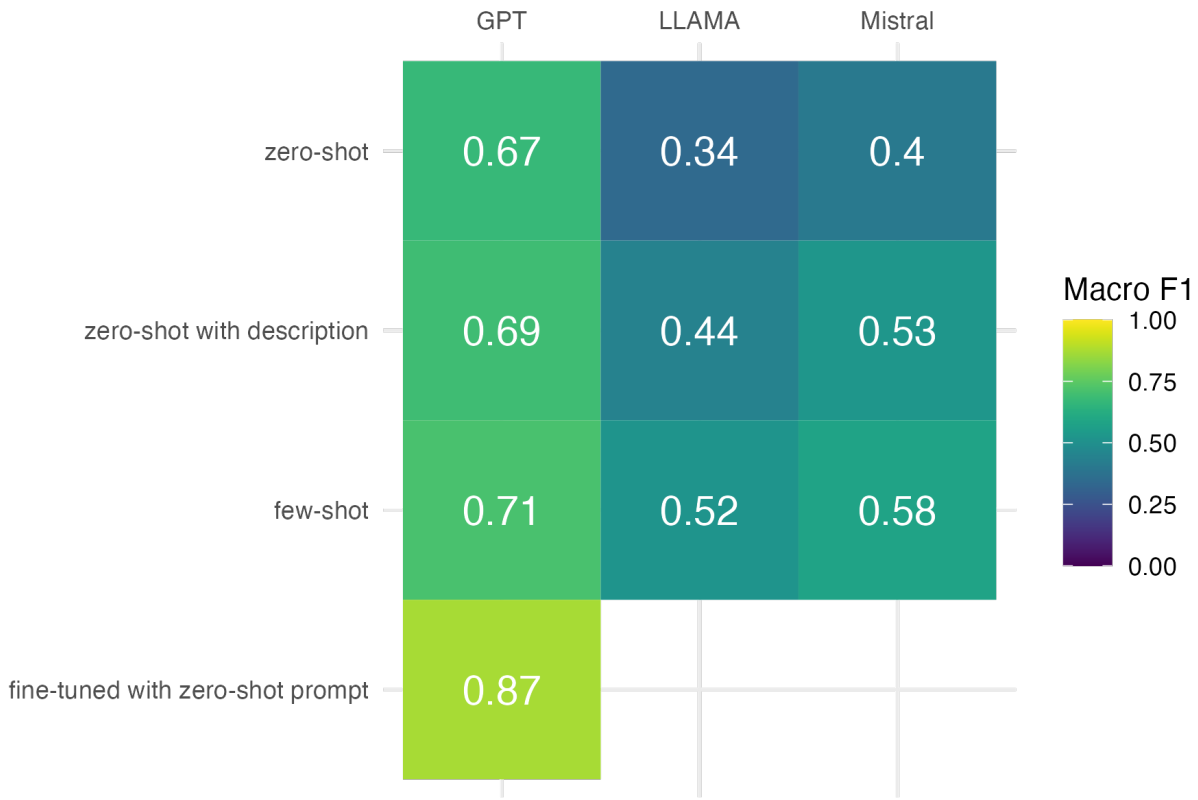


Figure 4.2: Macro F1 scores by LLM and prompting approach.

Distributions. The differences in LLMs' classification performance across categories result in different frequencies of categories (see Figure 4.4), although the overall shape of the distribution is similar to the human-coded benchmark. While the LLMs' good performance on classifying *incentive*, *interest*, and *fun* leads to the proportion of responses in these categories being close to the human benchmark, their poor performance on other categories manifests in substantially lower proportions than the human data would suggest. This includes *non-identifiable/other*, which is among the five most frequently identified categories according to the human coders. Llama and Mistral additionally assign too few cases to *no reason*, but code more responses as *curiosity* than both humans and GPT, where they also perform worse in terms of F1 scores. Conversely, the

[7]For "learning" under zero-shot prompting (with and without definitions) and for "recruiter" when prompting with descriptions, no macro F1 scores can be calculated for Llama, indicating that there were no true positives, no false positives (i.e., the LLM did not assign any of the responses to that category) for these categories. The same is true for Mistral for the category "no reason" under zero-shot prompting.

proportion of responses assigned to *tell opinion* tends to be lower when using Llama. In contrast, Mistral assigns disproportionately many responses to *tell opinion*, and, to a smaller degree, to *help society*, *help science*, and *recruitment* – the categories where GPT tends to outperform.

The proportion of missing (including ambiguous) assignments is (initially) higher for the open-source models than for GPT. Just as with performance and overall distribution, GPT is also less sensitive to prompting approaches than other LLMs when it comes to missing classifications, whereas the performance of Llama and Mistral depends on the prompting approach. Both open-source models eventually return better results than GPT when considering the amount of missing classifications. When using GPT, missing classifications occur almost exclusively for responses labeled as *no reason* by human coders (Figure A3.3), with over 60% of responses lacking a classification. In contrast, missing classifications are more evenly distributed across all categories when using Llama (which also misses assignments for close to 60% of *no reason* responses) or Mistral. This partly helps explain the poor classification performance for the *no reason* category; however, missing values cannot account for the poor performance on other categories (see Appendix 3.6 for full confusion matrices; and Appendix 3.4 for F1 scores when omitting missing values).

Reliability. Turning to reliability of the classifications, Mistral's output is identical across the two iterations, proving to be the only LLM tested where setting the temperature to zero and setting a seed actually results in the desired behavior – returning identical and therefore reliable output.[8] Yet, the other two LLMs also exhibit high reliability (ICC > 0.93, see Table 4.1). There are only minimal differences, with GPT being slightly more reliable than Llama.

Approach	**GPT-4o**	**Llama 3.2**	**Mistral NeMo**
zero-shot	0.99	0.95	1.00
with descriptions	0.99	0.94	1.00
few-shot	0.99	0.95	1.00
fine-tuned	0.99		

Table 4.1: ICC (two-way agreement) between two rounds of coding per LLM and prompting approach.

In sum, there are differences between LLMs in terms of performance and, to a lesser extent, reliability when coding German open-ended survey responses. Disregarding prompting approaches, using GPT results in higher classification performance than using Llama or Mistral, but performance under GPT is still subpar, both in absolute terms and relative to other methods (e.g., Haensch et al., 2022) when not using fine-tuning (see below). While all LLMs exhibit high reliability across iterations, Mistral has a slight edge, reproducing the exact same classifications.

[8] It is possible that the discrepancies in the other two LLMs are caused by the temperature not being implemented as zero by the LLMs (despite setting it as such), but a very small number, for mathematical reasons. Temperature is a normalization parameter for the LLM's underlying softmax function; setting it to zero would result in division by zero. We briefly discuss the implications of this in the next section.

Figure 4.3: Per-category F1 scores by LLM and prompting approach.

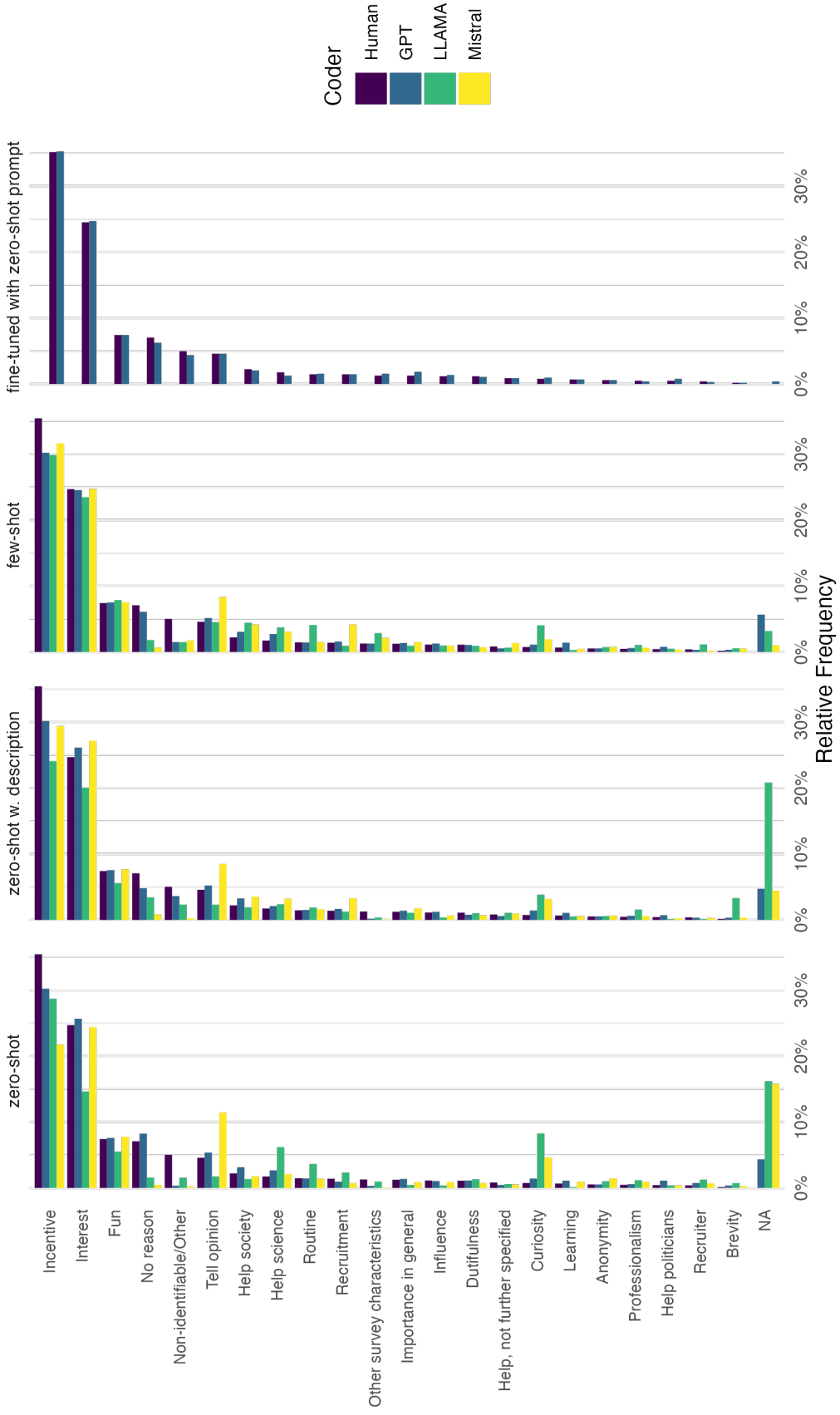


Figure 4.4: Distribution of coding categories by LLM and prompting approach.
n=5,072 for zero-shot (with and without description) and few-shot prompting, n=1,024 for fine-tuned prompting.

Differences between prompting approaches

Performance. When comparing differences in classification performance between prompting approaches across LLMs, performance is best for few-shot prompting and worst for zero-shot prompting in terms of macro F1. However, the size of the difference depends on the LLM used. There is a strong improvement in performance from zero-shot prompting to few-shot prompting when using the open-source models – for both models, there is a difference of 0.18 in macro F1 scores, see Figure 4.2. The same pattern emerges when considering other performance metrics (Appendix 3.5), and when investigating classification performance per category (Figure 4.3). However, for singular combinations of LLM used and category classified, performance is worse when providing the LLM with descriptions than when using simple zero-shot prompting (e.g., *fun*, *no reason*, *help science*), or when providing examples relative to providing descriptions (e.g., *recruitment*, *anonymity*, *non-identifiable/other*). This is more often so for the open-source LLMs than for GPT.

Most notably, GPT's performance drastically improves when employing fine-tuning, achieving a macro F1 of 0.87 – a 16-point difference over few-shot prompting and a satisfactory level in general. This jump can largely be attributed to much improved classification in the non-substantive categories. For other categories, a mixed picture emerges, with large improvements for six categories, but minor improvements for the remainder – in part because few-shot prompting already led to high levels of performance.

Distributions. Although all prompting approaches examined approximately result in very similar distributions of categories, few-shot prompting tends to approximate the distribution of the human-coded data best (Figure 4.4). This is especially the case for *interest* and *tell opinion*. Large differences remain especially for *non-identifiable/other*, *help science*, and *help society*. Few-shot prompting also results in substantially fewer responses that were not coded successfully, with a reduction of almost four fifths for Mistral. As a consequence, there are almost no missing classifications under few-shot prompting, except for *no reason* (Figure A3.3). Fine-tuning results in a distribution that perfectly matches the human classifications, with only four classifications missing in total (all belonging to the *no reason* category).

Reliability. All LLMs exhibit high reliability (>0.93) regardless of approach when considering ICCs (see Table 4.1). Mistral is completely deterministic in all approaches, GPT is consistently very reliable across approaches, including fine-tuning, and Llama is slightly less reliable when provided with descriptions.

To summarize, the prompting approach used does make a difference in terms of performance, but not so much in terms of reliability of coding German open-ended survey responses. Providing detailed descriptions of categories tends to improve classification performance over zero-shot prompting, and few-shot prompting further improves it, especially for the open-source LLMs. Fine-tuning leads to the best overall performance and the largest improvement compared to other prompting approaches when using GPT. Reliability is high regardless of the prompting approach used.

4.5 Discussion

In our study, we assessed the performance and reliability of three powerful, multilingual LLMs (GPT-4o, Llama 3.2, and Mistral NeMo), when classifying German open-ended survey responses on a specific and complex topic given a pre-defined coding scheme. We also investigated differences depending on the prompting approach used. Overall, performance differed greatly between LLMs, and only a fine-tuned LLM achieved satisfactory levels of predictive performance (macro F1 of 0.87). In general, GPT performed best, and, disregarding fine-tuning, few-shot prompting led to the second-best performance (macro F1 of 0.71 for GPT), echoing the findings of previous studies on English data on less specific topics (Halterman & Keith, 2024; Mellon et al., 2024). Performance differences between prompting approaches were conditional on the LLM used – the prompting approach was not as important when using GPT, but made a big difference for other LLMs, especially Mistral. While the LLMs correctly identified most of the responses belonging to the most frequently occuring (and most easily identifiable) reasons, they struggled with non-substantive catch-all categories. Limitations in performance in these categories may arise because human coders classified responses such as "don't know", "xxx", and blank responses as *no reason*. The LLMs often failed to categorize such data, instead treating it as if it contained no response. This is problematic for open-ended response classification more broadly. Responses belonging to such categories are quite common regardless of question topic, as many survey respondents lack the time or motivation to respond to open-ended questions, either giving non-substantive or nonsensical responses that practically correspond to item-nonresponse (Krosnick & Presser, 2010). In our case, LLMs' unequal classification performance across different categories of reasons for survey participation results in different categorical distributions when not using fine-tuning. Such discrepancies could also have consequences for further inferential analyses of the coded data. Thus, LLM-coded open-ended responses could paint a very different picture of the concept being measured by a survey item than human coding would.

Our study shows that using off-the-shelf (i.e., non-fine-tuned) LLMs is not necessarily superior to other computational methods for coding open-ended responses. Comparing our results to those of Haensch et al. (2022), who used an SVM on the same data, even few-shot performance proved to be below expectations when going beyond the most obvious and common categories (median F1 0.83 vs. 0.72 at best). This is at odds with Mellon et al. (2024) findings regarding English-language survey responses on a more common topic: Although that study also reported that GPT models were superior to Llama models, it also found that the LLMs, when provided with the full coding scheme including descriptions and examples for over 50 categories, were much better at classifying British responses to the commonly discussed "most important problem" question than established supervised approaches, including BERT and SVMs. Rytting et al. (2023) came to similar conclusions even for the by now outdated GPT-3 under few-shot prompting, albeit for a task with only three categories. It thus appears that the applicability of LLMs for coding open-ended responses depends not just on the LLM and prompting approach used, but also on the topic (in terms of specificity and categorical complexity) and possibly language of the responses.

However, as our findings show, LLMs have the potential to match or even outperform other methods when fine-tuned. Using the zero-shot prompt on the fine-tuned GPT achieved a macro F1 of 0.87 (median F1 0.88), with dramatic improvements for non-substantive responses. This resulted in perfectly matched distributions between human and LLM-coded responses and virtually no missing classifications. Although this confirms speculations in terms of improved effectiveness over off-the-shelf usage (Mellon et al., 2024), it does not yet fulfill the hopes of being a resource-efficient alternative to established methods. This is because fine-tuning LLMs requires

a sufficiently large set of human-coded benchmark data and more computational resources and expertise, similar to established methods, with which researchers are often more familiar. In addition, such established methods usually do not require payment, whereas proprietary LLMs (potentially requiring less programming expertise if providing user-friendly interfaces for fine-tuning) do. Additionally, this approach, as all others, relies on a pre-defined coding scheme, which may not readily exist for all open-ended questions practitioners might want to have classified.

While all three models we examined were very reliable in their classifications across two iterations, only Mistral showed the desired behavior of identical output when setting the model temperature to zero and setting a seed. The possibility that setting the temperature to the least probabilistic setting does not actually guarantee deterministic behavior can be unintuitive for survey researchers not familiar with LLMs in-depth, potentially risking a false sense of confidence. Yet, even the deviating LLMs in our study were more reliable than previous studies suggested (e.g., Heseltine & Clemm von Hohenberg, 2024), making resolvement by human coders (which, in the aforementioned study, did not exhibit higher agreement) obsolete. However, reproducibility over longer periods of time, e.g., for several survey waves featuring the same open-ended item, is not guaranteed when using non-local models, due to them being subject to change or deprecation. This highlights the need for regular validation with humans in the loop (see also Weber & Reichardt, 2023), even under high performance (which we only observed for the fine-tuned approach).

Our results also highlight the trade-offs between proprietary and open-source LLMs in terms of cost, privacy, reliability, and performance. Using open-source models such as Llama and Mistral, available on platforms such as Huggingface, are free to use and can be run locally, ensuring privacy and reproducibility by avoiding third-party servers and model updates. However, running them requires considerable computing resources and expertise, which not all researchers may have access to. In contrast, proprietary models like GPT, while user-friendly, incur costs per token (i.e., input and output length), which can be high for large datasets or complex instructions.[9] In our case, open-source LLMs underperformed compared to proprietary ones in coding open-ended responses, and fine-tuning a GPT model was the most successful approach. Finally, the speed of advancement of LLMs presents researchers with the challenge of working towards a moving target, where working with reliable and reproducible model versions may not present the state of the art.

Our work gives rise to some further considerations and possible improvements. First, more experiments with different prompting strategies (Schulhoff et al., 2025) could be explored to see whether fine-tuned performance can be neared or made more cost-effective. For example, even more explicit instructions emphasizing the importance of always assigning a category and exactly one category might improve results especially for non-substantive responses. Researchers could also investigate whether breaking down the task into a two-step process would reduce its complexity by shortening the coding scheme information to be processed per prompt, and lead to more satisfactory results. In this prompt-chaining approach, the LLM could first be asked whether a specific category would be suitable for an answer. After having iterated across all possible categories in the coding scheme, the LLM could then be asked for the best-suited category

[9] Per iteration through the dataset, we spent about EUR 10 for zero-shot prompting with GPT, EUR 20 for zero-shot prompting with descriptions, EUR 15 for few-shot prompting, and EUR 60 for fine-tuning and zero-shot prompting the fine-tuned model. Considering that in our case, inference took between 2 and 6 hours per iteration depending on the prompting approaches when self-hosting Llama 3 and Mistral on a A100 GPU, if renting such resources cost around EUR 2 per hour, this would result in an estimated cost of EUR 4-12 per iteration. Fine-tuning on the same dataset might require 4-6 hours, adding a one-time cost of about EUR 10-15. However, precise cost estimation is difficult due to variability in model size, hardware availability, batch optimization, and additional engineering overhead.

from among the set of those it identified as suitable. Such an approach would allow for more examples per category in the first step without negatively impacting the LLM's context capacity (see, e.g., Mellon et al., 2024), thereby possibly improving performance. For fine-tuning, future research should focus on systematic experiments to identify the minimum amount of human-coded data needed for effective performance, balancing resource efficiency with accuracy. Additionally, LLMs' inner workings, including how they process different languages relative to one another, are somewhat opaque and not always consistent (see, e.g., Zhang et al., 2023) – they might be better aligned to follow English instructions and coding schemes regardless of the language of the text to be classified. It is thus possible that LLMs perform better on non-English text classification when instructed in English, i.e., when only the survey response is in the native language. This would allow for simultaneous coding and translation of open-ended survey responses (Heseltine & Clemm von Hohenberg, 2024). Future research could investigate this by employing the English translation of our prompt.

Second, our study focused on the performance and reliability of LLM-coded open-ended survey responses, without investigating the impact of the method on the findings of substantive analyses. Replications of earlier substantive analyses that used more established classification methods with a fine-tuned LLM could complement our research. As part of such an analysis, taking into account uncertainty could shed light on whether distributional differences between LLM-based and human classifications are systematic. This could be done by analyzing the LLM's internal token probabilities (i.e., the probability with which the output is chosen), choosing the majority category after multiple iterations using an LLM's default temperature, or by directly asking the LLM for its certainty in a specific label (e.g., Tian et al., 2023). However, if human coders are inconsistent, models may be unfairly penalized, leading to deceptively low accuracy metrics. Even high inter-rater agreement (e.g., Cohen's kappa) can mask systematic errors made consistently by humans and mimicked by the model.

Finally, LLMs might detect patterns or nuances humans do not, especially when not constrained by a fixed coding scheme. Using LLMs for unsupervised approaches, such as topic modeling (e.g., Ornstein et al., 2024), could address this concern while also making the ex-ante development of coding schemes for new survey items obsolete (Mellon et al., 2024), further increasing efficiency compared to supervised methods. However, results from unsupervised approaches are challenging to evaluate due to the absence of ground truth labels and because the interpretations of discovered patterns are often subjective (Pham et al., 2024). In addition, even if humans are subjective, the large discrepancy between human and LLM-based codes in our study suggests the latter are systematically mistaken (see Fröhling et al., 2024, for a suggestion for diversifying LLM annotation). Depending on the complexity of the response data, it thus appears that off-the-shelf LLMs are not able to capture human reasoning as expressed in open-ended survey responses when not fine-tuned with human-coded benchmark data.

4.6 Conclusion

At a time when LLMs are revolutionizing survey research, there have been high hopes for their applicability to coding open-ended survey responses. Other studies have demonstrated singular LLMs' promising potential when tasked to code responses in comparatively easy contexts. However, we have shown that these findings do not necessarily generalize to other topically or linguistically more complex contexts: There is no one-size-fits-all kinds of open-ended response data regarding the LLM or prompting approach used. Even for the same data, using just any LLM for

coding does not work equally well, nor does it work automatically without humans in the loop. Instead, it requires careful prompt engineering or, even better, fine-tuning with data pre-coded by humans. When coding German open-ended responses on a very specific topic with a complex classification scheme, LLM performance is generally low and differs greatly between LLMs. In addition, differences in prompting approaches are conditional on the LLM used. Comparing GPT, Llama and Mistral, using a fine-tuned version of GPT resulted in the highest classification performance. When not fine-tuning, however, classification quality is low compared to other, "easier" application contexts (English-language responses to more common survey items) and other classification methods (supervised machine learning models). LLMs may thus be an effective and possibly efficient alternative in such easier settings, provided a pre-defined coding scheme exists – but success is not guaranteed. Our results indicate that the specific LLM and prompting approach to be used for coding open-ended responses needs to be thoroughly validated before deployment. For more difficult (con)texts, fine-tuning on human-coded data increases the chances of success. Thus, as of now, humans still need to be in the loop for the coding and analysis of open-ended survey responses.

References

Accelerate. (n.d.). Retrieved January 30, 2025, from https://huggingface.co/docs/accelerate/index

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2025). Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 15–36. https://doi.org/10.1609/icwsm.v19i1.35801

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 1–15. https://doi.org/10.1017/pan.2023.2

Barrie, C., Palmer, A., & Spirling, A. (2024). Replication for Language Models: Problems, Principles, and Best Practice for Political Science. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Bosley, M., Jacobs-Harukawa, M., Licht, H., & Hoyle, A. (2023). Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel. *Social Science Computer Review*, *36*(1), 103–115. https://doi.org/10.1177/0894439317697949

Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for Market Research. https://doi.org/10.2139/ssrn.4395751

Cerina, R., & Duch, R. (2023, September). Artificially Intelligent Opinion Polling [arXiv:2309.06029 [stat]]. Retrieved September 21, 2023, from http://arxiv.org/abs/2309.06029

Cuevas, A., Brown, E. M., Scurrell, J. V., Entenmann, J., & Daepp, M. I. G. (2023, October). Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models [arXiv:2309.10187 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2309.10187

da Costa-Luis, C., Larroque, S. K., Altendorf, K., Mary, H., richardsheridan, Korobov, M., Yorav-Raphael, N., Ivanov, I., Bargull, M., Rodrigues, N., Shawn, Dektyarev, M., Górny, M., mjstevens777, Pagel, M. D., Zugnoni, M., JC, CrazyPython, Newey, C., … McCracken, J. (2024, November). Tqdm: A fast, Extensible Progress Bar for Python and CLI. https://doi.org/10.5281/ZENODO.595120

Durmus, E., Nguyen, K., Liao, T., 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. (2024). Towards Measuring the Representation of Subjective Global Opinions in Language Models. *First Conference on Language Modeling*. https://openreview.net/forum?id=zl16jLb91v

Fröhling, L., Demartini, G., & Assenmacher, D. (2024, October). Personas with Attitudes: Controlling LLMs for Diverse Data Annotation [arXiv:2410.11745]. Retrieved October 23, 2024, from http://arxiv.org/abs/2410.11745

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). Irr: Various Coefficients of Interrater Reliability and Agreement. https://CRAN.R-project.org/package=irr

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–74. https://doi.org/10.1257/jel.20181020

GESIS. (2024). GESIS Panel - Extended Edition [Published: GESIS, Cologne. ZA5664 Data file Version 54.0.0, https://doi.org/10.4232/1.14385]. https://doi.org/10.4232/1.14385

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*. https://doi.org/10.1037/met0000540

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Gweon, H., & Schonlau, M. (2024). Automated Classification for Open-Ended Questions with BERT. *Journal of Survey Statistics and Methodology*, *12*(2), 493–504. https://doi.org/10.1093/jssam/smad015

Haensch, A.-C., Weiß, B., Steins, Patricia, Chyvra, Priscilla, & Bitz, Katja. (2022). The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis. *Frontiers in Big Data*, *5:880554*. https://doi.org/10.3389/fdata.2022.880554

Halterman, A., & Keith, K. A. (2024, July). Codebook LLMs: Adapting Political Science Codebooks for LLM Use and Adapting LLMs to Follow Codebooks [arXiv:2407.10747]. Retrieved October 11, 2024, from http://arxiv.org/abs/2407.10747

Hand, D. J., Christen, P., & Ziyad, S. (2024, September). Selecting a classification performance measure: Matching the measure to the problem [arXiv:2409.12391 [cs]]. Retrieved October 2, 2024, from http://arxiv.org/abs/2409.12391

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text [Publisher: SAGE Publications Ltd]. *Research & Politics*, *11*(1), 20531680241236239. https://doi.org/10.1177/20531680241236239

Hub client library. (n.d.). Retrieved January 30, 2025, from https://huggingface.co/docs/huggingface_hub/index

Joachims, T. (2001). A statistical learning model of text classification for support vector machines [event-place: New Orleans, Louisiana, USA]. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 128–136. https://doi.org/10.1145/383952.383974

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022, March). The Ghost in the Machine has an American accent: Value conflict in GPT-3 [arXiv:2203.07785 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2203.07785

Kim, J., & Lee, B. (2023, November). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction [arXiv:2305.09620 [cs]]. Retrieved January 23, 2024, from http://arxiv.org/abs/2305.09620

Krosnick, J., & Presser, S. (2010). Question and Questionnaire Design. In P. Marsden & J. Wright (Eds.), *Handbook of Survey Research. 2nd edition* (pp. 263–314). Emerald.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Landesvatter, C. (2024). *Methods for the classification of data from open-ended questions in surveys* [Doctoral dissertation, University of Mannheim]. https://madoc.bib.uni-mannheim.de/67089/

Lebrun, B., Temtsin, S., Vonasch, A., & Bartneck, C. (2024). Detecting the corruption of online questionnaires by artificial intelligence. *Frontiers in Robotics and AI, Volume 10 - 2023*. https://doi.org/10.3389/frobt.2023.1277635

Li, B., Haider, S., & Callison-Burch, C. (2024). This Land is Your, My Land: Evaluating Geopolitical Bias in Language Models through Territorial Disputes. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3855–3871. https://doi.org/10.18653/v1/2024.naacl-long.213

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023, September). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve [arXiv:2309.13638 [cs]]. Retrieved September 18, 2024, from http://arxiv.org/abs/2309.13638

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. v. d. Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Meidinger, M., & Aßenmacher, M. (2021). A New Benchmark for NLP in Social Sciences: Evaluating the Usefulness of Pre-trained Language Models for Classifying Open-ended Survey Responses: *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, 866–873. https://doi.org/10.5220/0010255108660873

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, *11*(1). https://doi.org/10.1177/20531680241231468

Meta. (2024a, September). Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Retrieved January 23, 2025, from https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

Meta. (2024b, December). Llama-3.2-3B-Instruct. Retrieved December 20, 2024, from https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

MistralAI. (2024a). Mistral-Nemo-Instruct-2407. Retrieved December 20, 2024, from https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

MistralAI. (2024b, July). Mistral NeMo [Section: news]. Retrieved December 20, 2024, from https://mistral.ai/news/mistral-nemo/

Ooi, H., httr development team, Littlefield, T., Holden, S., Stone, C., & Microsoft. (2019, February). AzureAuth: Authentication Services for Azure Active Directory [Institution: Comprehensive R Archive Network Pages: 1.3.3]. Retrieved January 30, 2025, from 10.32614/CRAN.package.AzureAuth

OpenAI. (2024a). Hello GPT-4o. Retrieved December 19, 2024, from https://openai.com/index/hello-gpt-4o/

OpenAI. (2024b, August). *GPT-4o System Card* (tech. rep.). Retrieved December 19, 2024, from https://cdn.openai.com/gpt-4o-system-card.pdf

Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2024). How to Train Your Stochastic Parrot: Large Language Models for Political Texts. https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf

Palmer, A., Smith, N. A., & Spirling, A. (2023). Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, *4*(1), 2–3. https://doi.org/10.1038/s43588-023-00585-1

Pangakis, N., Wolken, S., & Fasching, N. (2023, May). Automated Annotation with Generative AI Requires Validation [arXiv:2306.00176 [cs]]. https://doi.org/10.48550/arXiv.2306.00176

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M.,

Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pezeshkpour, P., & Hruschka, E. (2024, June). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006–2017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.130

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2024, June). TopicGPT: A Prompt-based Topic Modeling Framework. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2956–2984). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.164

Porst, R., & von Briel, C. (1995, May). *Wären Sie vielleicht bereit, sich gegebenenfalls noch einmal befragen zu lassen? Oder: Gründe für die Teilnahme an Panelbefragungen.* (tech. rep. No. 95/04). Zentrum für Umfragen, Methoden und Analysen. Mannheim. https://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/zuma_arbeitsberichte/95_04.pdf

R Core Team. (2024, July). R: The R Project for Statistical Computing. Retrieved July 2, 2024, from https://www.r-project.org/

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023, June). Towards Coding Social Science Datasets with Language Models [arXiv:2306.02177 [cs]]. Retrieved October 17, 2023, from http://arxiv.org/abs/2306.02177

Schierholz, M., & Schonlau, M. (2021). Machine Learning for Occupation Coding—A Comparison Study. *Journal of Survey Statistics and Methodology*, *9*(5), 1013–1034. https://doi.org/10.1093/jssam/smaa023

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions [Artwork Size: 143-152 Pages Publisher: European Survey Research Association]. *Survey Research Methods*, *Vol 10*, 143–152 Pages. https://doi.org/10.18148/SRM/2016.V10I2.6213

Schonlau, M., Weiß, J., & Marquardt, J. (2023). Multi-label classification of open-ended questions with BERT. *2023 Big Data Meets Survey Science (BigSurv)*, 1–8. https://doi.org/10.1109/BigSurv59479.2023.10486634

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2025). The Prompt Report: A Systematic Survey of Prompt Engineering Techniques [_eprint: 2406.06608]. https://arxiv.org/abs/2406.06608

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science [Bandiera_abtest: a Cg_type: World View Publisher: Nature Publishing Group Subject_term: Ethics, Machine learning, Technology, Scientific community]. *Nature*, *616*(7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. (2023, December). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In H. Bouamor, J. Pino,

& K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5433–5442). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.330

Törnberg, P. (2024). Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages [Publisher: SAGE Publications Inc]. *Social Science Computer Review*, 08944393241286471. https://doi.org/10.1177/08944393241286471

von der Heyde, L., Haensch, A.-C., Wenz, A., & Ma, B. (2024). United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections [Version Number: 2]. https://doi.org/10.48550/ARXIV.2409.09045

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-t., Tu, Z., & Lyu, M. (2024, August). Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6349–6384). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.345

Weber, M., & Reichardt, M. (2023, December). Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models [arXiv:2401.00284]. Retrieved October 24, 2024, from http://arxiv.org/abs/2401.00284

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 [Number: 43]. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020, October). Transformers: State-of-the-Art Natural Language Processing. https://doi.org/10.5281/zenodo.7391177

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023, December). Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7915–7927). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.491

5 Discussion and Conclusion

Summary and discussion of findings

The goal of this dissertation was to investigate the conditions under which LLMs can be applied in survey research to ensure high data quality. In the introduction, I provided a theoretical overview of the potential applications of LLMs in the survey research process and the data quality challenges such applications might bring with them. In the three substantive chapters of this dissertation, I focused on two major applications of LLMs, simulating respondents and coding open-ended responses, covering both representational and measurement challenges in LLM-based survey research. I tested these applications in previously unexamined contexts – European countries and languages – that are likely more challenging for LLMs than previous successful applications in the United States. This way, I both widened the societal and linguistic scope of applying LLMs in survey research, and provided empirical evidence for some of the previously identified potential biases of this tool, which limit the generalizability of its applicability. In this concluding chapter, I will summarize and discuss the results of the presented studies more broadly and in light of current developments and highlight some avenues for future research before providing some concluding remarks.

I refer to Chapter 1 as well as the discussion sections of the individual studies for more detailed summaries of their research designs and findings. Here, I will focus on the overarching conclusions that can be drawn from them, and their implications. Most importantly, the studies show that a major shortcoming of LLMs in the context of survey research is their **lack of nuance** in reflecting human attitudes and behavior in the examined low-resource contexts. This holds both for predicting voting behavior and for classifying reasons for survey participation. Regarding voting behavior, LLMs' predictions differed from both survey-reported (Chapter 2) and actual (Chapter 3) behavior, both on the individual (Chapter 2) and aggregate (Chapters 2 and 3) level, and regarding both turnout (Chapter 3) and party choice (Chapters 2 and 3). They were biased towards Green and Left parties and their voters (Chapters 2 and 3), partisans, and other "typical" voters, missing the complex factors that determine vote choice. Moreover, LLMs' predictions failed especially for Eastern European countries with Slavic native languages (Chapter 3). Taken together, this evidence supports the argument that LLMs tend to simplify and generalize across contexts they have less information on – compared to, for example, voting behavior in the United States. Regarding reasons for survey participation, LLM's classifications were accurate for more common and obvious categories, but less so for less common and non-substantive catch-all categories. Ultimately, as discussed in Chapter 4, the observed lack of nuance results in different distributions and correlative relationships of the measured concepts, risking erroneous substantive conclusions.

This lack of nuance in mirroring humans is contrasted by a **need for detail** to arrive at somewhat acceptable results. Once again, this was evident in both applications. In Chapter 3, I showed that predictions of voting behavior worsened when given only socio-demographic information about voters. In Chapter 4, I demonstrated that few-shot prompting, i.e., providing LLMs

with examples of desired output, led to better classification performance than various forms of zero-shot prompting without such examples. It should thus not be surprising that fine-tuning an LLM with information pertinent to the task at hand further improves results (e.g., Gururangan et al., 2020; Wei et al., 2022). The research I presented in this dissertation only tested and confirmed this approach for the application of coding open-ended responses (Chapter 4). Research building on the study presented in Chapter 2 also suggests fine-tuning as a promising approach for estimating public opinion with LLM-based synthetic samples (see below for a more detailed discussion of Holtdirk et al., 2024). The implications of these findings are twofold. On the one hand, they present a promising outlook – LLMs might be able to mirror human attitudes and behavior, if given enough detail through re-adjusting its weights. On the other hand, current, off-the-shelf LLMs apparently are not the resource-efficient, easily accessible tool survey researchers and practitioners have hoped them to be. As I discuss in Chapters 3 and 4, recent, detailed, target-population specific (survey) data is often not available, and fine-tuning requires computational expertise and resources. In addition, if LLMs' predictions are conditional on the information researchers deem relevant enough to include in a prompt, they are based on information researchers already have. This is reminiscent of the "people machine" in President Kennedy's campaign, which was deemed one of the first examples of artificial intelligence, but did not tell the campaign anything it did not already know about voting groups. The missing-data problem only gets transferred to LLMs. Off-the-shelf, general-purpose LLMs thus, at least at the time of writing, have very limited added value over established survey research tools.

The studies featured in this dissertation also highlight the **differences between proprietary and open-source LLMs**. In both tested applications, models of the proprietary GPT family performed better than open-source models of the Llama and Mistral families (Chapters 3 and 4). As I discuss extensively in Chapter 4, researchers face considerable trade-offs when deciding between open- and closed-source LLMs regarding performance, resources (both computational and financial), expertise, privacy, and replicability. From a research ethics perspective, there are strong arguments for using open-source LLMs (Barrie et al., 2024; Palmer et al., 2023; Spirling, 2023). In this light, research such as that I presented here is especially informative: I have shown that the applicability of open-source LLMs differs across survey research tasks – while they performed worse than GPT in both tasks, the differences were much smaller in the coding of open-ended responses (Chapter 4) than in predicting voting behavior Chapter 3. Furthermore, my research in Chapter 4 shows that prompting approaches are differentially effective across LLMs – for GPT, the approach used did not matter much, whereas it made a large difference for Mistral. All in all, these results show that there is no "one-size-fits-all" approach to using LLMs for survey research tasks – the choices of LLM and prompting approach are interrelated and depend on the task.

When discussing differences between LLM versions, temporal factors need to be considered as well. This dissertation covers research spanning two years, conducted in a consecutive order. During this time, LLMs continued to develop rapidly. The studies presented in this dissertation always featured the most recent, most powerful models of their respective developers at the respective time of writing. They therefore allow insights into the **improvements of LLMs over time** – for example, the output of the Mistral LLM used in Chapter 4 was much more concise than that used in Chapter 3, which resulted in less missing data and therefore more positive results. Such observations give rise to optimism – if the trend continues, we might likely see LLMs be more applicable in the survey research process, producing more accurate, less error-prone results. Especially considering the need for reproducible research, the research community might be encouraged to see that the performance of open-source LLMs appears to be catching up with

that of their proprietary counterparts. The recent launch of DeepSeek (DeepSeek-AI et al., 2025), a powerful open-source LLM being on par with OpenAI's latest models' performance at much lower cost (Gibney, 2025), is a promising step in this direction. Of course, these observations also imply that the conclusions drawn in this dissertation might not be definitive and stand the test of time. Conversely, one can conclude that LLMs are likely to further improve, and previous results have to constantly be re-evaluated in light of new technological developments. In addition, these observations underline the importance of research (such as that presented in this dissertation) pointing out the shortcomings of existing models, so that active steps towards improvement can be taken.

Overall, the results of the studies presented in this dissertation highlight the **context-dependency of the applicability of LLMs** in the survey research process. LLM-based synthetic samples are not equally applicable for estimating the attitudes and behaviors of global populations, at least not when relying on off-the-shelf LLMs (Chapters 2 and 3). Even more so, their need for recent and detailed information means they for now have little added value over survey-based measures, as I discussed in Chapter 3. Similarly, using off-the-shelf LLMs for coding language- and topic-specific open-ended responses does not yield results that are comparable with human coding (Chapter 4). The need for fine-tuning to achieve satisfactory results implies that, at the time of writing, there is little advantage over established semi-automated approaches. In summary, the applicability of LLMs for survey-related tasks not only depends on the LLM and prompting approach used, but also on the context of the task: In low-resource contexts, LLM-based approaches are less likely to succeed. Importantly, "low-resource" can relate to the language (non-English), the task (e.g., predicting human behavior or classifying human text), the specificity of the topic (e.g., vote choice, survey motivation), and the complexity (number of potential categories). The biases identified in this dissertation showcase the disparities between high- and low-resource contexts in LLMs – in this case, the disparities in their internal representation of global and individual diversity of human attitudes and behavior. Finally, the previous chapters have shown that LLMs are more suitable for some applications in the survey research process (e.g., coding open-ended responses) than others (e.g., simulating respondents), adding another dimension to the context-dependency of the applicability of LLMs for survey research.

Current trends and future research

The findings of the preceding chapters also give rise to further research regarding the identification and mitigation of biases in LLMs in the context of survey research. Limitations and further research related to the specific use cases have been addressed in the individual chapters. I will now outline some more general aspects of ongoing research and opportunities for future research. This includes others' selected research that extends the studies presented in this dissertation, which first were released as preprints.

The cases selected in the previous chapters for testing the limitations of LLM-applicability for survey research in low-resource contexts could be considered comparatively easy – European linguistic and societal contexts are still mostly WEIRD. As I have argued before, however, the research presented here can also be informative beyond these test cases, for even-lower-resource contexts. Nevertheless, **explicit tests** are needed. For example, Qi et al. (2024) recently confirmed the arguments put forth in Chapters 2 and 3: Comparing persona-based estimations of U.S., German, and Chinese voting behavior simulated by GPT-3.5-Turbo to representative election studies, they found that performance was better in English-speaking countries and two-party

systems. Future research using more recent LLMs could provide insights into which biases remain across time (LLM families and versions) and space (task and population contexts). Similarly, Qu and Wang (2024) recently found that persona-based simulations generated with ChatGPT more closely matched the political attitudes of English-speaking countries (South Africa, Singapore), particularly the U.S., than countries with "smaller" native languages (Brazil, Sweden, Japan) and that the LLM exhibited biases towards demographic subgroups.

The research designs featured in this dissertation have concentrated on factors that can be associated with biased LLM training data. However, as I have indicated earlier, **pinpointing the mechanisms causing** LLMs' outputs about human attitudes and behavior to be biased is challenging: As discussed in Chapters 2 and 3, even for high-resource contexts, biases have been observed. Likely, biases are due to a combination of lacking diversity in training data and alignment processes and model architectures – the major components behind any LLM output (McCoy et al., 2023). This "machine bias" might be idiosyncratic for each LLM (Boelaert et al., 2024). Due to the complexity and opacity of (black box) LLMs' inner workings, identifying their biases through their outputs, as I have done in the preceding chapters, is a necessary proxy for social scientists. The recently emerging reasoning models might not only perform better at solving more complex (survey research) tasks, they could also be prompted to elicit explicit reasoning steps for their output, thereby gaining further insights into their inner workings. Experimenting with different training data corpora would be an alternative approach for identifying the exact sources of bias. However, this would require extensive computational resources – the kind necessary for training an entire LLM, several times – as well as access to an LLMs' source code and weights, neither of which most social scientists have. In absence of such computing capacities, it is a viable option to investigate LLMs' latent space, that is, the layers of vectors containing the semantic representations based on which LLMs perform next-token predictions – at least for interpretable open-source LLMs. Replicating the data source and prompt design introduced in Chapter 2, Ball et al. (2025) investigate how LLMs map human attributes to party preferences in this latent space. Their findings echo those presented in this dissertation and of, e.g., Bisbee et al. (2024): Responses from instruction-tuned LLMs show a bias towards left-leaning parties. Overall, LLM-based responses exhibit less variance between demographic subgroups, but higher entropy within – indicating that LLM responses are subject to some artificially injected randomness, thereby failing to capture distinct subgroup-specific preferences. Opening "white box" LLMs thus confirms the findings of the research discussed in this dissertation that are based on evaluating output at face value. In addition, concurring with the observations of Perez et al. (2023) and Rozado (2024), Ball et al. (2025) find that base models are less left-leaning (in fact, more right-leaning) than instruction-tuned ones. This suggests that right-wing bias is introduced through Internet data, whereas left-leaning bias is introduced in the alignment processes, for example, through reinforcement learning with human feedback (RLHF) that, among others, aims at making output less harmful. This calls for a critical investigation of the belief systems that are encoded in LLMs not just indirectly through selective training data, but directly through the comparatively small and homogeneous group of human crowd workers performing RLHF tasks (Hovy & Prabhumoye, 2021; Kirk, Vidgen, et al., 2024), as they have the potential to mitigate as well as amplify the biases encoded in LLMs.[1]

[1] This does not imply that RLHF should be abolished. Instead, debiasing efforts should include a diversification of the perspectives that get encoded into LLMs through RLHF – i.e., ensuring diverse representation in the annotators (but see Rystrøm, 2023). Ultimately, while *who counts* is an empirical question, *who decides who should count* is a normative one. See Ferrara (2023), Kirk, Vidgen, et al. (2024), and Ryan et al. (2024) for further discussions, and Kirk, Whitefield, et al. (2024) for an example of a more diversified alignment dataset.

These observations about bias introduced through RLHF also challenge the survey research tradition's emphasis on human-generated data and whether such data should always be regarded as superior (see, e.g., Ornstein et al., 2024). For example, and in apparent irony to the considerations above, crowd workers employed for coding extensive (social science) datasets have begun using LLMs to do the work for them (Veselovsky et al., 2023), complicating the question of what researchers are comparing their LLM-based experiments to (Demszky et al., 2023).[2] At the same time, such considerations can serve as counter-arguments against using LLMs to detect patterns which human-generated data does not capture – a limitation I discussed in the individual chapters of this dissertation. As I argued in Chapter 1, LLMs have as much potential for introducing error in survey research as they have for mitigating it – sometimes, in the same task.

As I have argued throughout the preceding chapters, *general*-purpose LLMs were not designed for the *specific* purpose of any type of survey research application, neither through their training nor architecture. Therefore, developing LLMs customized for survey research would be a potential remedy. As a first step in this direction, researchers have begun experimenting with **fine-tuning** existing off-the-shelf LLMs (e.g., L. Li et al., 2024; Lin, 2024, see also Chapter 4). The training data used could consist of survey research best practices to aid with tasks related to survey design and implementation (e.g., Wenz & Haensch, 2024). For using LLMs as survey respondents, training LLMs on survey datasets could help align LLM output to that of specific populations (e.g., Cao et al., 2025; Kim and Lee, 2023, see also Simmons and Hare, 2023, whereas social media data could update their knowledge about current events and public reaction to those events (e.g., Ahnert et al., 2025; Chu et al., 2023). For example, extending the research featured in this dissertation, Holtdirk et al. (2024) fine-tuned the open-source LLMs featured in Chapters 3 and 4 for predicting German vote choice, relying on the same data source and prompt variables as the study presented in Chapter 2. Their results indicate that fine-tuning balances out the overestimation of voters of left-leaning parties. Replicating the voting prediction study by Argyle et al. (2023), they also show that the fine-tuning approach can be transferred to other contexts. Of course, the selected prompt variables cannot capture all the determinants of human opinion formation. For even more detailed information, in-depth interviews could be used to tune LLM agents that answer respondents' questions for them (Park et al., 2024); however, this only appears feasible on a smaller scale. On a larger scale, national survey data archives could collaborate to generate an immense dataset of the attitudes and behaviors of global populations, including longitudinal, cross-sectional, and cross-national information, which could then be used as the training dataset for a more general public opinion-LLM (Bail, 2024; Kreuter, 2025). This process would make the need for fine-tuning LLMs anew for every new survey context obsolete. Additionally, as the findings of this dissertation indicate, LLMs need better alignment to specific target groups in terms of underrepresented languages and cultural values, for example through specific (re-)training or through multi-LLM collaboration (Ali et al., 2024; Feng et al., 2024; C. Li et al., 2024; NLLB Team et al., 2024; Ostendorff et al., 2024; TrustLLM, n.d.). Such efforts should be accompanied by collaborative research teams bringing together technical experts and members of the public, especially the communities impacted by misaligned LLMs (Bail, 2024; D'Ignazio and Klein, 2020; Hovy and Prabhumoye, 2021; see also Anthropic, 2023 for an example of democratizing AI alignment.

Ultimately, there is a need for more transparency in LLM design, regarding training data and alignment processes as well as model architectures (e.g., Hardinges et al., 2024). Working with

[2] This also has consequences for the re-training of LLMs with survey-specific data – if that data is actually generated by LLMs, the LLMs are being trained on their own output, gradually removing human diversity from their knowledge, which, as some argue, risks eventually leading to model collapse (Shumailov et al., 2024).

open-source LLMs therefore is not only justified by their performative advancement and advantages in fine-tuning outlined above, but also warranted by the need for alignment, interpretability, and replicability (Bail, 2024; Palmer et al., 2023; Senoner et al., 2024; Spirling, 2023).

Besides fine-tuning, **prompt engineering** is another important aspect to investigate when it comes to optimizing LLMs for survey research – after all, prompts are the part of LLM input researchers as end users have the most control over. As I have shown in Chapters 3 and 4, varying the *amount and type of information* contained in a prompt can have substantial impact on output quality. Future research could shed light onto whether more context-specific information in the prompt, related to, e.g., determinants of voting behavior for the synthetic sampling use case, is sufficient for customizing LLMs to specific populations, or whether fine-tuning is always necessary. As I have argued earlier, however, both alternatives rely on survey data, limiting the advantage of LLMs. In addition, modifications in *order and wording of the information* might change outputs. Findings regarding this aspect are contested, with some studies identifying profound effects (e.g., Pezeshkpour & Hruschka, 2024), others none (e.g., Hartmann et al., 2023; Moore et al., 2024). For example, in the context of Chapter 2 of this dissertation, Ball et al. (2025) found that some LLMs were very sensitive to paraphrasing of the persona prompts, while others were not. However, the work by Wang et al. (2024) indicates that this might depend on whether the output probabilities or the text output is examined. These competing findings call for more research and underline the importance of LLM-specific prompt design.

The large-scale proliferation of LLMs has hit society and research like a meteor, which will continue to experience the shock waves of its impact. Filling the research crater it left while it continues to expand with every new model release requires ongoing, interdisciplinary work. A research agenda for LLMs in survey research, and in the social sciences more broadly, could stand at the outset of this endeavor. Such an agenda would identify which pressing ethical, methodological, and substantive questions of LLM use in empirical social science research have been addressed and are yet to be addressed. As I mentioned in its introduction, this dissertation had the aim of contributing to the effort of filling this crater. Going forward, gaps related to scope as well as methodology need to be addressed. I focused on two of the major applications of LLMs in survey research – LLMs acting as respondents and as research assistants for text analysis. However, as I detailed in Chapter 1, there are many more potential use cases of LLMs in the survey research process. There is a need for both a systematic theoretical review of existing and systematic empirical evaluation of untested LLM applications across a range of surveys and populations. Ideally, such an evaluation would be carried out along a unified framework that allows researchers to quantify biases and have specific standards for acceptable performance. Knowing which practices amplify and mitigate biases would allow researchers to minimize them in their research design and safeguard data quality, thereby ensuring valid inferences for research, policymaking, and society as a whole. Such error frameworks have proven successful for survey and digital trace data, but the novelty and idiosyncrasy of LLMs calls for yet another **adaptation of the Total Survey Error (TSE) framework** (Groves & Lyberg, 2010; Groves et al., 2009) to the LLM-augmented reality of survey research. Several of the error sources previously identified by Pennell et al. (2017), Amaya et al. (2020), and Sen et al. (2021) for multinational surveys, Big Data, and digital trace data can likely be transferred to LLM-assisted survey research, but, as outlined in Chapter 1 LLMs' idiosyncratic features also introduce new error sources. LLMs are a tool with many screws, such as model choice, hyperparameters, or prompt design. These screws could be mapped to different parts of the TSE. Integrating traditional, previously identified, and LLM-specific errors into a unified framework will be a helpful contribution to both the survey research community and the computer science and natural language processing (NLP) community that is developing

LLMs, which would be provided with guidance for identifying biases, contributing to efforts to mitigate them. As another service to the survey research community, developing an overview of approaches and best practices for ensuring data quality for different kinds of LLM applications in survey research is a task for future work. For these efforts, it would be valuable to engage with the computer science and NLP community, which has been working on understanding and improving LLMs from a technical point of view (e.g., Gallegos et al., 2024; Hovy & Prabhumoye, 2021). When finding a shared vocabulary (see, e.g., Simmons & Hare, 2023), social science and computer science fields can create synergies for improving data quality for and of LLMs.

Conclusion

While much remains to be explored regarding the use of LLMs in empirical social research, this dissertation offered some valuable contributions. It showcased the potentials of LLMs in survey research, but, more importantly, provided evidence for what is not possible as of now – using off-the-shelf LLMs for reflecting human attitudes in "low-resource" contexts – (sub)populations and tasks that are not adequately represented in LLMs' training, alignment, and architecture. More broadly, this dissertation therefore also is informative for the development and de-biasing efforts of LLMs, both for survey research and other research areas more generally.

As I have mentioned repeatedly throughout this dissertation, LLMs are designed for predicting the most likely next word in a sentence in general – not for that sentence to represent public opinion. Succeeding in doing the latter – or any specific task – depends on LLMs' input, that is, their training and alignment, and the used prompt. I have argued and demonstrated that, as a result, the applicability of LLMs for survey research is context-dependent: both input aspects need to match the respective target population in order for output to be accurate. For better or for worse, this implies that survey researchers should undertake a fitness-for-purpose assessment of LLMs for the specific task at hand. The research discussed in this dissertation has shown that LLMs cannot fully replace humans in survey research, neither as respondents, researchers, nor research assistants. They can, however, augment human survey research with proper supervision and validation to prevent harm (see also Bail, 2024; Demszky et al., 2023; Jansen et al., 2023; Sarstedt et al., 2024). Thus, just as has been the case in light of other technological advancements, the work of survey researchers does not simply disappear in the age of AI – it shifts. Here, how the field responded to past technological advancements offers opportunities for learning for the present. Survey research has succeeded in integrating a variety of methods into its toolbox in the past – now, LLMs are added to the pile. I follow the thoughts expressed by Couper (2013, 2024) in the context of survey research in changing technological landscapes: amid all the noise caused by LLMs (literally and figuratively), survey researchers and computational social scientists must not lose focus on the people. They need to widen their gaze and consider integrations of different technologies in their research more broadly, rather than the specifics of one technology. The methodologies of using these technologies need to continually be synthesized and standardized while the latter keep developing. This includes continuing to be thorough in research designs and evaluations, being aware of what is and is not possible (and for what purpose), and developing new standards incorporating both the discipline's foundations and new potentials and pitfalls. As I have demonstrated in this dissertation, LLMs can mirror (only) broad patterns of human attitudes and behavior. With the necessary knowledge about their limitations and errors, it is possible that these broad patterns could be integrated with more precise and representative measurement tools to provide a better picture of how societies think and act. For this effort to succeed,

continued research on how to identify, adjust for, and ultimately, mitigate LLM biases is needed. In conclusion, even if machines are performing the counting for survey research, it is humans who are responsible for ensuring that *who counts* are all human voices.

References

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2025). Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 15–36. https://doi.org/10.1609/icwsm.v19i1.35801

Ali, M., Fromm, M., Thellmann, K., Ebert, J., Weber, A. A., Rutmann, R., Jain, C., Lübbering, M., Steinigen, D., Leveling, J., Klug, K., Buschhoff, J. S., Jurkschat, L., Abdelwahab, H., Stein, B. J., Sylla, K.-H., Denisov, P., Brandizzi, N., Saleem, Q., ... Flores-Herr, N. (2024, October). Teuken-7B-Base & Teuken-7B-Instruct: Towards European LLMs [arXiv:2410.03730 [cs]]. https://doi.org/10.48550/arXiv.2410.03730

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Anthropic. (2023, October). Collective Constitutional AI: Aligning a Language Model with Public Input. Retrieved February 26, 2025, from https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 1–15. https://doi.org/10.1017/pan.2023.2

Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Ball, S., Allmendinger, S., Kreuter, F., & Kühl, N. (2025, February). Human Preferences in Large Language Model Latent Space: A Technical Analysis on the Reliability of Synthetic Data in Voting Outcome Prediction [arXiv:2502.16280 [cs]]. https://doi.org/10.48550/arXiv.2502.16280

Barrie, C., Palmer, A., & Spirling, A. (2024). Replication for Language Models: Problems, Principles, and Best Practice for Political Science. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

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*, 1–16. https://doi.org/10.1017/pan.2024.5

Boelaert, J., Coavoux, S., Ollion, E., Petev, I. D., & Präg, P. (2024, April). Machine Bias: How do Generative Language Models Answer Opinion Polls? https://doi.org/10.31235/osf.io/r2pnb

Cao, Y., Liu, H., Arora, A., Augenstein, I., Röttger, P., & Hershcovich, D. (2025, February). Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations [arXiv:2502.07068 [cs]]. https://doi.org/10.48550/arXiv.2502.07068

Chu, E., Andreas, J., Ansolabehere, S., & Roy, D. (2023, March). Language Models Trained on Media Diets Can Predict Public Opinion [arXiv:2303.16779 [cs]]. https://doi.org/10.48550/arXiv.2303.16779

Couper, M. P. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, *7*(3), 145–156. https://doi.org/https://doi.org/10.18148/srm/2013.v7i3.5751

Couper, M. P. (2024, March). New Data Types and Surveys: Opportunities and Challenges. Retrieved April 2, 2025, from https://www.upf.edu/documents/244683118/246905697/

Couper_New+Data+Types+and+Surveys.pdf/4e6e367c-a12d-d2aa-a386-d8910a22f12c?t=1713766229952

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025, January). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [arXiv:2501.12948 [cs]]. https://doi.org/10.48550/arXiv.2501.12948

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00241-5

D'Ignazio, C., & Klein, L. F. (2020, March). *Data Feminism* [_eprint: https://direct.mit.edu/book-pdf/2390355/book_9780262358521.pdf]. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., & Tsvetkov, Y. (2024, November). Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4151–4171). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.240

Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models [arXiv:2304.03738 [cs]]. *First Monday*. https://doi.org/10.5210/fm.v28i11.13346

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Gibney, E. (2025). Scientists flock to DeepSeek: How they're using the blockbuster AI model [Bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Machine learning]. *Nature*. https://doi.org/10.1038/d41586-025-00275-0

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. John Wiley & Sons.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020, July). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8342–8360). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Hardinges, J., Simperl, E., & Shadbolt, N. (2024). We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models [Publisher: The MIT Press]. *Harvard Data Science Review*, (Special Issue 5). https://doi.org/10.1162/99608f92.a50ec6e6

Hartmann, J., Schwenzow, J., & Witte, M. (2023, January). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation [arXiv:2301.01768 [cs]]. Retrieved March 28, 2023, from http://arxiv.org/abs/2301.01768

Holtdirk, T., Assenmacher, D., Bleier, A., & Wagner, C. (2024, October). Fine-Tuning Large Language Models to Simulate German Voting Behaviour (Working Paper). https://doi.org/10.31219/osf.io/udz28

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12432]. *Language and Linguistics Compass*, *15*(8), e12432. https://doi.org/10.1111/lnc3.12432

Jansen, B. J., Jung, S.-g., & Salminen, J. (2023). Employing large language models in survey research. *Natural Language Processing Journal*, *4*, 100020. https://doi.org/10.1016/j.nlp.2023.100020

Kim, J., & Lee, B. (2023, November). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction [arXiv:2305.09620 [cs]]. Retrieved January 23, 2024, from http://arxiv.org/abs/2305.09620

Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, *6*(4), 383–392. https://doi.org/10.1038/s42256-024-00820-y

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024). The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models.

Kreuter, F. (2025). Modernizing Data Collection. *Journal of Official Statistics*, *41*(3), 863–872. https://doi.org/10.1177/0282423X251318452

Li, C., Teney, D., Yang, L., Wen, Q., Xie, X., & Wang, J. (2024, November). CulturePark: Boosting Cross-cultural Understanding in Large Language Models [arXiv:2405.15145 [cs] version: 3]. https://doi.org/10.48550/arXiv.2405.15145

Li, L., Li, J., Chen, C., Gui, F., Yang, H., Yu, C., Wang, Z., Cai, J., Zhou, J. A., Shen, B., Qian, A., Chen, W., Xue, Z., Sun, L., He, L., Chen, H., Ding, K., Du, Z., Mu, F., … Dong, Y. (2024, December). Political-LLM: Large Language Models in Political Science [arXiv:2412.06864 [cs]]. https://doi.org/10.48550/arXiv.2412.06864

Lin, H. (2024, December). Designing Domain-Specific Large Language Models: The Critical Role of Fine-Tuning in Public Opinion Simulation [arXiv:2409.19308 [cs] version: 2]. https://doi.org/10.48550/arXiv.2409.19308

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023, September). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve [arXiv:2309.13638 [cs]]. Retrieved September 18, 2024, from http://arxiv.org/abs/2309.13638

Moore, J., Deshpande, T., & Yang, D. (2024, July). Are Large Language Models Consistent over Value-laden Questions? [arXiv:2407.02996 [cs]]. Retrieved July 30, 2024, from http://arxiv.org/abs/2407.02996

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., … Wang, J. (2024). Scaling neural machine translation to 200 languages. *Nature*, *630*(8018), 841–846. https://doi.org/10.1038/s41586-024-07335-x

Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2024). How to Train Your Stochastic Parrot: Large Language Models for Political Texts. https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf

Ostendorff, M., Suarez, P. O., Lage, L. F., & Rehm, G. (2024). LLM-Datasets: An Open Framework for Pretraining Datasets of Large Language Models.

Palmer, A., Smith, N. A., & Spirling, A. (2023). Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, *4*(1), 2–3. https://doi.org/10.1038/s43588-023-00585-1

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024, November). Generative Agent Simulations of 1,000 People [arXiv:2411.10109]. Retrieved November 18, 2024, from http://arxiv.org/abs/2411.10109

Pennell, B.-E., Hibben, K. C., Lyberg, L. E., Mohler, P. P., & Worku, G. (2017, February). A Total Survey Error Perspective on Surveys in Multinational, Multiregional, and Multicultural Contexts. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total Survey Error in Practice* (pp. 179–201). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119041702.ch9

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2023, July). Discovering Language Model Behaviors with Model-Written Evaluations. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13387–13434). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.847

Pezeshkpour, P., & Hruschka, E. (2024, June). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006–2017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.130

Qi, W., Lyu, H., & Luo, J. (2024, July). Representation Bias in Political Sample Simulations with Large Language Models [arXiv:2407.11409 [cs] version: 1]. https://doi.org/10.48550/arXiv.2407.11409

Qu, Y., & Wang, J. (2024). Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, *11*(1), 1095. https://doi.org/10.1057/s41599-024-03609-x

Rozado, D. (2024). The political preferences of LLMs (T. Zhang, Ed.). *PLOS ONE*, *19*(7), e0306621. https://doi.org/10.1371/journal.pone.0306621

Ryan, M. J., Held, W., & Yang, D. (2024, August). Unintended Impacts of LLM Alignment on Global Representation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16121–16140). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.853

Rystrøm, J. H. (2023, June). Apolitical Intelligence? Auditing Delphi's responses on controversial political issues in the US [arXiv:2306.13000 [cs]]. https://doi.org/10.48550/arXiv.2306.13000

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21982]. *Psychology & Marketing*, *41*(6), 1254–1270. https://doi.org/10.1002/mar.21982

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A TOTAL ERROR FRAMEWORK FOR DIGITAL TRACES OF HUMAN BEHAVIOR ON ONLINE PLATFORMS. *Public Opinion Quarterly*, *85*, 399–422. https://doi.org/10.1093/poq/nfab018

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human–AI collaboration. *Scientific Reports*, *14*(1), 31150. https://doi.org/10.1038/s41598-024-82501-9

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Simmons, G., & Hare, C. (2023). Large Language Models as Subpopulation Representative Models: A Review [https://arxiv.org/abs/2310.17888]. https://doi.org/10.48550/arXiv.2310.17888

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science [Bandiera_abtest: a Cg_type: World View Publisher: Nature Publishing Group Subject_term: Ethics, Machine learning, Technology, Scientific community]. *Nature*, *616*(7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4

TrustLLM. (n.d.). TrustLLM: Democratizing Trustworthy and Factual Large Language Model Technology for Europe. Retrieved August 21, 2024, from https://trustllm.eu/

Veselovsky, V., Ribeiro, M. H., & West, R. (2023, June). Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks [arXiv:2306.07899 [cs]]. https://doi.org/10.48550/arXiv.2306.07899

Wang, X., Hu, C., Ma, B., Röttger, P., & Plank, B. (2024, August). Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think [arXiv:2404.08382 [cs]]. https://doi.org/10.48550/arXiv.2404.08382

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned Language Models are Zero-Shot Learners. *International Conference on Learning Representations*. https://openreview.net/forum?id=gEZrGCozdqR

Wenz, A., & Haensch, A.-C. (2024). Using large language models for evaluating and improving survey questions. *Proceedings of the 26th General Online Research Conference*. https://www.gor.de/wp-content/uploads/2024/02/ConferenceProceedings_2024_final.pdf

Appendices

A1 Appendix to Chapter 2

A1.1 GLES Questionnaire and GPT Prompt Values

The GLES is based on a multi-stage, stratified, random sample drawn from population registers in Germany (GLES n.d.). Survey participants are interviewed in computer-assisted personal interviews (CAPI).
For details about the primary dataset, see GLES (2019).

Variable	**GLES Questionnaire (German)**	**GLES Questionnaire (Translation by GLES)**	**GPT Prompt Values (German, as prompted) [Translation by authors]**
Age	Würden Sie mir bitte sagen, in welchem Jahr Sie geboren wurden?	What year were you born in?	2017 - year of birth
Gender	Intervieweranweisung: Ist die Zielperson männlich oder weiblich? (1) Männlich (2) Weiblich	Interviewer instruction: Is the respondent male or female? (1) Male (2) Female	**männlich** [male] if Gender = 1 **weiblich** [female] if Gender = 2
Education	*Schulabschluss* Welchen höchsten allgemeinbildenden Schulabschluss haben Sie? (1) Schule beendet ohne Abschluss (2) Hauptschulabschluss, Volksschulabschluss, Abschluss der polytechnischen Oberschule 8. oder 9. Klasse (3) Realschulabschluss, Mittlere Reife, Fachschulreife oder Abschluss der polytechnischen Oberschule 10.	*School leaving certificate* What's your highest level of general education? (1) Finished school without school leaving certificate (2) Lowest formal qualification of Germany's tripartite secondary school system, after 8 or 9 years of schooling (3) Intermediary secondary qualification, after 10 years of schooling (4) Certificate fulfilling entrance requirements to study at a polytechnical	**keinen Schulabschluss** [no degree] if School leaving certificate = 1 | 9 **einen Hauptschulabschluss** [Hauptschule degree] if School leaving certificate = 2 **einen Realschulabschluss** [Realschule degree] if School leaving certificate = 3 | 6 **Abitur** [Abitur degree] if School leaving certificate = 4 | 5

	Klasse (4) Fachhochschulreife (Abschluss einer Fachoberschule etc.) (5) Abitur bzw. erweiterte Oberschule mit Abschluss 12. Klasse (Hochschulreife) (6) Anderen Schulabschluss, und zwar: (9) Bin noch Schüler *Berufliche Bildung* Und welchen beruflichen Ausbildungsabschluss haben Sie? Nennen Sie mir bitte den Kennbuchstaben für den auf Sie zutreffenden Ausbildungsabschluss. (A) D - Beruflich-betriebliche Anlernzeit mit Abschlusszeugnis, aber keine Lehre (B) G - Teilfacharbeiterabschluss (C) I - Abgeschlossene gewerbliche oder landwirtschaftliche Lehre (D) B - Abgeschlossene kaufmännische Lehre (E) E - Berufliches Praktikum, Volontariat (F) M - Berufsfachschulabsc	college (5) Higher qualification, entitling holders to study at a university (6) Other school leaving certificate, please enter: (9) Still at school *Vocational and professional training* And what kind of vocational training did you complete? Please name the appropriate letter which corresponds with your vocational training. (A) D - On-the-job vocational training with final certificate, but not within a traineeship or apprenticeship scheme (B) G - Compact vocational training course (C) I - Completed trades/crafts or agricultural traineeship (D) B - Completed commercial traineeship (E) E - Work placement/internship (F) M - Specialized vocational college certificate (G) A - Vocational academy certificate (H) P - Technical or vocational college certificate	**einen Hochschulabschluss** [College degree] if Vocational and professional training = J | K | L | M [letter K, N, O, L]

	hluss (G) A - Fachakademie-/ Berufsakademieabschluss (H) P - Fachschulabschluss (I) H - Meister, Technikerabschluss (J) K - Fachhochschulabschluss (K) N - Hochschulabschluss: Bachelor (L) O - Hochschulabschluss: Master, Diplom, Magister, Staatsexamen (M) L - Promotion (N) C - Anderen Beruflichen Ausbildungsabschluss, und zwar: (O) F - Noch in beruflicher Ausbildung (P) J - Keine abgeschlossene Ausbildung	("Fachschulabschluss") (I) H - Master (craftsman), technician or equivalent college certificate (J) K - Polytechnic degree (K) N - University degree, Bachelor (L) O - University degree, Master (M) L - Doctoral degree (N) C - Other vocational training certificate, please enter: (O) F - Still training/studying (P) J - No completed vocational training	
Net household income	Wie hoch ist das monatliche Netto-Einkommen IHRES HAUSHALTES INSGESAMT? Ich meine dabei die Summe, die nach Abzug von Steuern und Sozialversicherungsbeiträgen übrig bleibt. Bitte ordnen Sie Ihr Haushaltseinkommen in die Kategorien der Liste ein und nennen	Taken all together, would you please indicate what the monthly net income of your household is? By net income, I mean the amount that you have left after taxes and social security. Please select the monthly net income of your household from one of these groups and tell me the group letter.	**niedriges** [low] if Net household income = 1 | 2 | 3 | 4 | 5 **mittleres** [medium] if Net household income = 6 | 7 | 8 | 9 | 10 **hohes** [high] if Net household income = 11 | 12 | 13

	Sie mir den Buchstaben. (1) B - unter 500 Euro (2) T - 500 bis unter 750 Euro (3) P - 750 bis unter 1000 Euro (4) F - 1000 bis unter 1250 Euro (5) E - 1250 bis unter 1500 Euro (6) H - 1500 bis unter 2000 Euro (7) L - 2000 bis unter 2500 Euro (8) N - 2500 bis unter 3000 Euro (9) R - 3000 bis unter 4000 Euro (10) M - 4000 bis unter 5000 Euro (11) S - 5000 bis unter 7500 Euro (12) A - 7500 bis unter 10000 Euro (13) D - 10000 Euro und mehr	(1) B - Less than 500 euros (2) T - 500 to less than 750 euros (3) P - 750 to less than 1000 euros (4) F - 1000 to less than 1250 euros (5) E - 1250 to less than 1500 euros (6) H - 1500 to less than 2000 euros (7) L - 2000 to less than 2500 euros (8) N - 2500 to less than 3000 euros (9) R - 3000 to less than 4000 euros (10) M - 4000 to less than 5000 euros (11) S - 5000 to less than 7500 euros (12) A - 7500 to less than 10000 euros (13) D - 10000 euros or more	
Employment status	Nun weiter mit der Erwerbstätigkeit und Ihrem Beruf. Was von dieser Liste trifft auf Sie zu? (1) Vollzeit berufstätig (mehr als 30 Stunden/Woche) (2) Teilzeit berufstätig (bis 30 Stunden/ Woche) (3) Lehrling/Azubi (4) Schüler (5) Student (6) In Umschulung	Do you currently work in a full-time or part-time job? Which of the descriptions in this list describes your status? (1) In full-time employment (more than 30 h/week) (2) In part-time employment (up to 30 h/week) (3) In a traineeship or apprenticeship	**nicht berufstätig** [not working] if Employment status = 7 | 10 | 12 **in Ausbildung** [studying/training] if Employment status = 3 | 4 | 5 | 6 | 9 **berufstätig** [working] if Employment status = 1 | 2 | 8 | 11

	(7) Zurzeit arbeitslos (8) Zurzeit in Kurzarbeit (9) Bundesfreiwilligendienst, Freiwilliges Soziales Jahr (FSJ), Freiwilliges Ökologisches Jahr (FÖJ) (10) Pensionär/Rentner (früher voll berufstätig) (11) In Mutterschutz/Elternzeit (12) Nicht berufstätig (Hausfrau/ Hausmann)	scheme (4) School student (5) Studying at a polytechnic or university (6) Currently on a retraining course (7) Currently unemployed (8) Currently on short-time working (9) Community service (Bundesfreiwilligendienst, Freiwilliges Soziales Jahr (FSJ), Freiwilliges Ökologisches Jahr (FÖJ)) (10) Retirement, on a pension (formerly employed) (11) On maternity leave, parental leave (12) Not in full or part-time employment (Housewife/Household)	
Religiosity	Was würden Sie von sich sagen? Sind Sie überhaupt nicht religiös, nicht sehr religiös, etwas religiös oder sehr religiös? (1) Überhaupt nicht religiös (2) Nicht sehr religiös (3) Etwas religiös (4) Sehr religiös	What would you say about yourself, are you not religious at all, not very religious, somewhat religious or very religious? (1) Not religious at all (2) Not very religious (3) Somewhat religious (4) Very religious	**überhaupt nicht religiös** [not at all religious] if Religiosity = 1 **nicht sehr religiös** [not very religious] if Religiosity = 2 **etwas religiös** [somewhat religious] if Religiosity = 3 **sehr religiös** [very religious] if Religiosity = 4
Left-right self-placement	Und wie ist das mit Ihnen selbst? Wo würden Sie sich auf	Where would you place yourself on this scale?	**stark links** [strongly left] if Left-right self-placement = 1 | 2

	der Skala von 1 bis 11 einordnen? (1) 1 links (2) 2 (3) 3 (4) 4 (5) 5 (6) 6 (7) 7 (8) 8 (9) 9 (10) 10 (11) 11 rechts	(1) 1 Left (2) 2 (3) 3 (4) 4 (5) 5 (6) 6 (7) 7 (8) 8 (9) 9 (10) 10 (11) 11 Right	**mittig links** [center-left] if Left-right self-placement = 3 | 4 **in der Mitte** [in the middle] if Left-right self-placement = 5 | 6 | 7 **mittig rechts** [center-right] if Left-right self-placement = 8 | 9 **stark rechts** [strongly right] if Left-right self-placement = 10 | 11
Party identification	Und nun noch einmal kurz zu den politischen Parteien. In Deutschland neigen viele Leute längere Zeit einer bestimmten politischen Partei zu, obwohl sie auch ab und zu eine andere Partei wählen. Wie ist das bei Ihnen: Neigen Sie - ganz allgemein gesprochen - einer bestimmten Partei zu? Und wenn ja, welcher? [Liste für Interviewer] (1) CDU/CSU (2) CDU (3) CSU (4) SPD (5) FDP (6) GRÜNE (7) DIE LINKE (322) AfD (801) andere Partei, und zwar (808) keine Partei	Now, let's look at the political parties. In Germany, many people lean towards a particular party for a long time, although they may occasionally vote for a different party. How about you, do you in general lean towards a particular party? If so, which one? [List for interviewer] (1) CDU/CSU (2) CDU (3) CSU (4) SPD (5) FDP (6) GRÜNE (7) DIE LINKE (322) AfD (801) [other party: specify] (808) [no party]	**mit der Partei CDU/CSU** [CDU/CSU] if Party identification = 1 | 2 | 3 **mit der Partei SPD** [SPD] if Party identification = 4 **mit der Partei FDP** [FDP] if Party identification = 5 **mit der Partei Bündnis 90/Die Grünen** [Greens] if Party identification = 6 **mit der Partei Die Linke** [Left] if Party identification = 7 **mit der Partei AfD** [AfD] if Party identification = 322 **mit einer Kleinpartei** [small/other party] if Party identification = 801 **mit keiner Partei** [not with any party] if Party identification = 808

Strength of party identification	Wie stark oder wie schwach neigen Sie - alles zusammengenommen - dieser Partei zu: sehr stark, ziemlich stark, mäßig, ziemlich schwach oder sehr schwach? (1) Sehr stark (2) Ziemlich stark (3) Mäßig (4) Ziemlich schwach (5) Sehr schwach	All in all, how strongly or weakly do you lean toward this party: very strongly, fairly strongly, moderately, fairly weakly or very weakly? (1) Very strongly (2) Fairly strongly (3) Moderately (4) Fairly weakly (5) Very weakly	**sehr schwach** [very weakly] if Strength of party identification = 5 **ziemlich schwach** [rather weakly] if Strength of party identification = 4 **mäßig** [moderately] if Strength of party identification = 3 **ziemlich stark** [rather strongly] if Strength of party identification = 2 **sehr stark** [very strongly] if Strength of party identification = 1
East/West Germany	(0) Ostdeutschland (1) Westdeutschland	[coded by interviewer] 0 East Germany 1 West Germany	**Westdeutschland** [West Germany] 0 **Ostdeutschland** [East Germany] 1
Immigration	Und wie ist Ihre Position zum Thema Zuzugsmöglichkeiten für Ausländer? Bitte benutzen Sie diese Skala. (1) 1 Zuzugsmöglichkeiten für Ausländer sollten erleichtert werden (2) 2 (3) 3 (4) 4 (5) 5 (6) 6 (7) 7 (8) 8 (9) 9 (10) 10 (11) 11 Zuzugsmöglichkeiten für Ausländer sollten	And what position do you take on immigration for foreigners? Please use the scale. (1) 1 Immigration for foreigners should be easier (2) 2 (3) 3 (4) 4 (5) 5 (6) 6 (7) 7 (8) 8 (9) 9 (10) 10 (11) 11 Immigration for foreigners should be more difficult	**einschränken** [limit] if Immigration = 7 | 8 | 9 | 10 | 11 **weder erleichtern noch einschränken** [neither nor] if Immigration = 6 **erleichtern** [facilitate] if Immigration = 1 | 2 | 3 | 4 | 5

	eingeschränkt werden		
Inequality	Es gibt zu verschiedenen politischen Themen unterschiedliche Meinungen. Wie ist das bei Ihnen: Was halten Sie von folgenden Aussagen? Bitte antworten Sie anhand der Liste. (D) Die Regierung sollte Maßnahmen ergreifen, um die Einkommensunterschiede zu verringern. (1) Stimme voll und ganz zu (2) Stimme eher zu (3) Teils/teils (4) Stimme eher nicht zu (5) Stimme überhaupt nicht zu	There are various opinions on different political issues. What do you think of the following statements? Please use the list. (D) The government should take measures to reduce the differences in income levels. (1) Strongly agree (2) Agree (3) Neither agree nor disagree (4) Disagree (5) Strongly disagree	**keine Maßnahmen ergreifen** [don't take measures] if Inequality= 4 | 5 **habe keine Meinung dazu, ob die Regierung Maßnahmen ergreifen sollte** [no opinion] if Inequality= 3 **Maßnahmen ergreifen** [take measures] if Inequality= 1 | 2

Table A1.1: GLES variables and corresponding prompt variables.

A1.2 Example Prompts

1. Main sample

English (translation)	**German (as prompted)**
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]	Ich bin **28** Jahre alt und **weiblich**. Ich habe **einen Hochschulabschluss**, ein **mittleres** monatliches Haushalts-Nettoeinkommen und bin **berufstätig**. Ich bin **nicht religiös**. Politisch-ideologisch ordne ich mich **mittig links** ein. Ich identifiziere mich **ziemlich schwach** mit der Partei **Bündnis 90/Die Grünen**. Ich lebe in **Westdeutschland**. Ich finde, die Regierung sollte die Einwanderung **erleichtern** und **Maßnahmen ergreifen**, um die Einkommensunterschiede zu verringern. Habe ich bei der Bundestagswahl 2017 gewählt und wenn ja, welcher Partei habe ich meine Zweitstimme gegeben? Ich habe [INSERT]

Table A1.2a: Example prompt for the main study and English translation (variables bolded).

Note: We decided not to include "gewählt" (voted) as a suffix in the prompt, using the [MASK] instead of [INSERT] request, as it might bias the output against non-voters by reducing the likelihood of GPT completing the sentence with "nicht" (not) or "ungültig" (invalid) due to German semantics. We leave the further exploration of these effects to prompt engineering researchers.

2. Robustness sample (respondents with missing values)

Any sentence except the first one was omitted if the respective variable was missing for the respondent.

English (translation)	**German (as prompted)**
I am **28** years old. I am **female**. I have a **college** degree. I have a **medium** monthly net household income. I 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. I think the government should **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]	Ich bin **28** Jahre alt. Ich bin weiblich. Ich habe **einen Hochschulabschluss**. Ich habe ein **mittleres** monatliches Haushalts-Nettoeinkommen. Ich bin **berufstätig**. Ich bin **nicht religiös**. Politisch-ideologisch ordne ich mich **mittig links** ein. Ich identifiziere mich **ziemlich schwach** mit der Partei **Bündnis 90/Die Grünen**. Ich lebe in **Westdeutschland**. Ich finde, die Regierung sollte die Einwanderung **erleichtern**. Ich finde, die Regierung sollte **Maßnahmen ergreifen**, um die Einkommensunterschiede zu verringern. Habe ich bei der Bundestagswahl 2017 gewählt und wenn ja, welcher Partei habe ich meine Zweitstimme gegeben? Ich habe [INSERT]

Table A1.2b: Example prompt for robustness checks and English translation (variables bolded).

A1.3 Model Configuration Parameters and Pilot

According to OpenAI, the difference between the chat completions-API (the API accessing the model embedded in ChatGPT and therefore commonly referred to as “ChatGPT”) and the completions-API lies in the underlying models and cost, with the chat completions-API offering access to the more capable GPT-4 and cost-effective GPT-3.5-turbo, which corresponds to the performance of the completion-API’s text-davinci-003, which we used, at a fraction of the cost (OpenAI, 2022). Text-davinci-003 is a version of GPT-3.5 optimized for efficient text completion (OpenAI, 2024b).

We opted for specifying the randomness (temperature) rather than restricting the completion-sample to the tokens above a certain probability threshold (parameter top-p), as recommended by OpenAI. We further opted not to specify a penalty for the prompt information provided on party identification and/or ideology, leaving the details of this for further research in prompt engineering. The following configurations were tested sequentially on a subsample of five personas:

1. Max. Tokens

Specifies the maximum amount of tokens a completion may contain. One token corresponds to about four letters in the English language.
Default values for other parameters: temperature = 1, n=1

- Test 1: maxtoken = 20: 1/3 did not contain party; 20 tokens used per completion
- Test 2: maxtoken = 30: works as desired, 1 "insert" but contains party; 10-25 tokens used per completion
- Test 3: maxtoken = 40: one contradictory and one ambiguous case, but complete information; 20-30 tokens used per completion
- Test 4: maxtoken = 50: complete information; 30-40 tokens used per completion

30 tokens provide ample space for GPT to complete the prompt with a full sentence containing vote choice. This does not imply that all of GPT’s completions will be 30 tokens in length. Indeed, our analyses reveal that a substantial portion of completions break off despite using less than 30 tokens.

→ **DECISION ON TOKENS: 30**

2. Temperature

Specifies the randomness of possible completions by specifying the sampling strategy from the underlying distribution. 0 corresponds to no randomness (deterministic), with repetitive completions, 1 corresponds to complete randomness with maximum variation. Default in OpenAI Playground: 1. For detailed explanations, see Argyle et al. (2023).
Values for other parameters: maxtokens = 30, n=1

- Test 5: temperature = 0.9 (rgpt3 package default): complete information, 1/3 incomplete sentence, 25-30 tokens used per completion
- Test 6: temperature = 0.7 (Argyle et al. 2023): complete information, complete sentences, 25-30 tokens used per completion. Differences in person with ambiguous predictors

→ **DECISION ON TEMPERATURE: 0.9**

3. Multiple completions vs. Best of

N specifies the number of completions per prompt (default: 1). *Best of* determines the space of possibilities from which to select the completion with the highest probability. Generates best_of completions server-side and returns the "best" (the one with the highest log probability per token).
Values for other parameters: maxtokens = 30, temperature = 0.9

- Test 7: n = 5:
    - Records 5 completions
    - Returns warning: To avoid an `invalid_request_error`, `best_of` was set to equal `n`
- Test 8: best_of = 5 Records only 1 completion

→ **DECISION ON N vs BEST OF: N = 5, best_of = 1 to account for inability to store token probabilities.**
Function will force value to default to n.

A1.4 Accepted completions for party matching

Small parties that were listed on the ballot for the 2017 election but were not represented in the 18th German Bundestag (2013-2017) are summarized as “Small party”.

Party / GLES reported vote **(**Question q19ba; [brackets]: translation)	**GPT completion contains** (case-insensitive; *asterisk*: embedded within any word; **bold**: conceptually equivalent wordings)
CDU/CSU	CDU, CSU, CDU/CSU, Union, *christ*
SPD	SPD, *sozialdemokrat*
Bündnis 90/Die Grünen [Greens]	*Grün*, 90, Bündnis
FDP	FDP, freie, *liberal*
Die Linke [Left]	*link* [confirmed by manual check]
AfD	AfD, Alternative [confirmed by manual check]
Andere Partei [other / small party]	Andere [confirmed by manual check] **Kleinpartei** [confirmed by manual check] ***any small party names, e.g., Piraten*** [confirmed by manual check]
Ungültig gewählt [invalid vote]	[confirmed by manual check] ungültig **keine Zweitstimme**
Nicht gewählt [did not vote]	[confirmed by manual check] nicht, **keine Partei, weder gewählt noch eine Zweitstimme abgegeben**

Table A1.3: Keywords used for matching between GPT completions and German political parties / GLES vote choice.

A1.5 Documentation of manual checks

- If the completion contained “a left party” without specification for “Die Linke”, it was recoded as NA and resampled
- Manual checks were performed if multiple party names were extracted and/or “Erststimme” (primary vote) was mentioned, in order to extract the correct party
    - If multiple parties were named and no distinction between Erst- and Zweitstimme was made, the second-named party was assumed to be the Zweitstimme
    - If “Erststimme” and “Zweitstimme” were mentioned and it was not clear which vote the party name referred to, it was recoded as NA
    - If “Erststimme” but no “Zweitstimme” was mentioned, it was recoded as NA
    - If the completion contained “meine Stimme” (my vote) and only one party without an indication of whether this was Erst- or Zweitstimme, Zweitstimme was assumed and the vote recorded accordingly
- Manual checks were performed for all completions that couldn’t be matched with any vote choice automatically
- Hallucinations were recoded as NAs and resampled. Reasoning: face-validity (no party on the ballot in Germany or perfectly matchable to one party)
- If the completion contained the equivalent of “voted” followed by “did not vote”, it was recoded as “did not vote”, as it became evident that completing the sentence “Ich habe” with “gewählt” was a common/meaningless default before elaborating on the choice made.
- Notes:
    - 1 completion contained “Parteienalternative” which was thus matched with AfD, but could also have been matched with “Other party”
    - 1 completion contained a party that cannot be further named [breakoff], which was wrongly matched as “did not vote” but could have been an NA

	Sample 1	**Sample 2**	**Sample 3**
Total completions	9525	1427	281
Total modified	653 (6.9%)	107 (7.5%)	27 (9.6%)
NAs (after modification)	1427 (14.9%)	281 (19.7%)	89 (31.7%)
Robustness: Total completions	1885	461	235
Robustness: Total modified	78 (4.1%)	20 (4.3%)	10 (4.3%)
Robustness: NAs (after modification)	461 (24.5%)	235 (51%)	141 (60%)

Table A1.4: Completions obtained from GPT, modifications and missing completions in main and robustness samples.

A1.6 Summary Statistics, GLES Data (Main Sample)

Variable	N	Mean / Prop.	Std. Dev.	Min.	Median	Max.
Age	1905	51	18	18	52	95
Gender	1905					
male	1001	53%				
female	904	47%				
Education	1905	2.5	1.2	0	2	4
Employment Status	1905					
not working	671	35%				
studying	143	8%				
working	1091	57%				
Income	1905					
low	349	18%				
medium	1324	70%				
high	232	12%				
Residence	1905					
West Germany	1289	68%				
East Germany	616	32%				
Religiosity	1905					
not at all	668	35%				
not very	320	17%				
somewhat	689	36%				
very	228	12%				
LR-Ideology	1905	-0.3	0.81	-2	0	2
Party ID	1905					
CDU/CSU	540	28%				
SPD	356	19%				
Greens	185	10%				
FDP	98	5%				
Left	159	8%				
AfD	83	4%				
Small party	23	1%				
none	461	24%				

Strength of Party ID	1905	2.7	1.7	0	3	5
Att. Inequality	1905					
act	1464	77%				
no opinion	251	13%				
don't act	190	10%				
Att. Immigration	1905					
facilitate	603	32%				
neither nor	350	18%				
limit	952	50%				
Vote Choice (GLES)	1905					
CDU/CSU	504	26%				
SPD	338	18%				
Greens	224	12%				
FDP	200	10%				
Left	188	10%				
AfD	162	9%				
Small party	73	4%				
Invalid vote	11	1%				
No vote	205	11%				
Imputed	1905					
no	1528	80%				
yes	377	20%				

Table A1.5: Summary statistics of prompt input variables, main sample.

A1.7 Model Performance

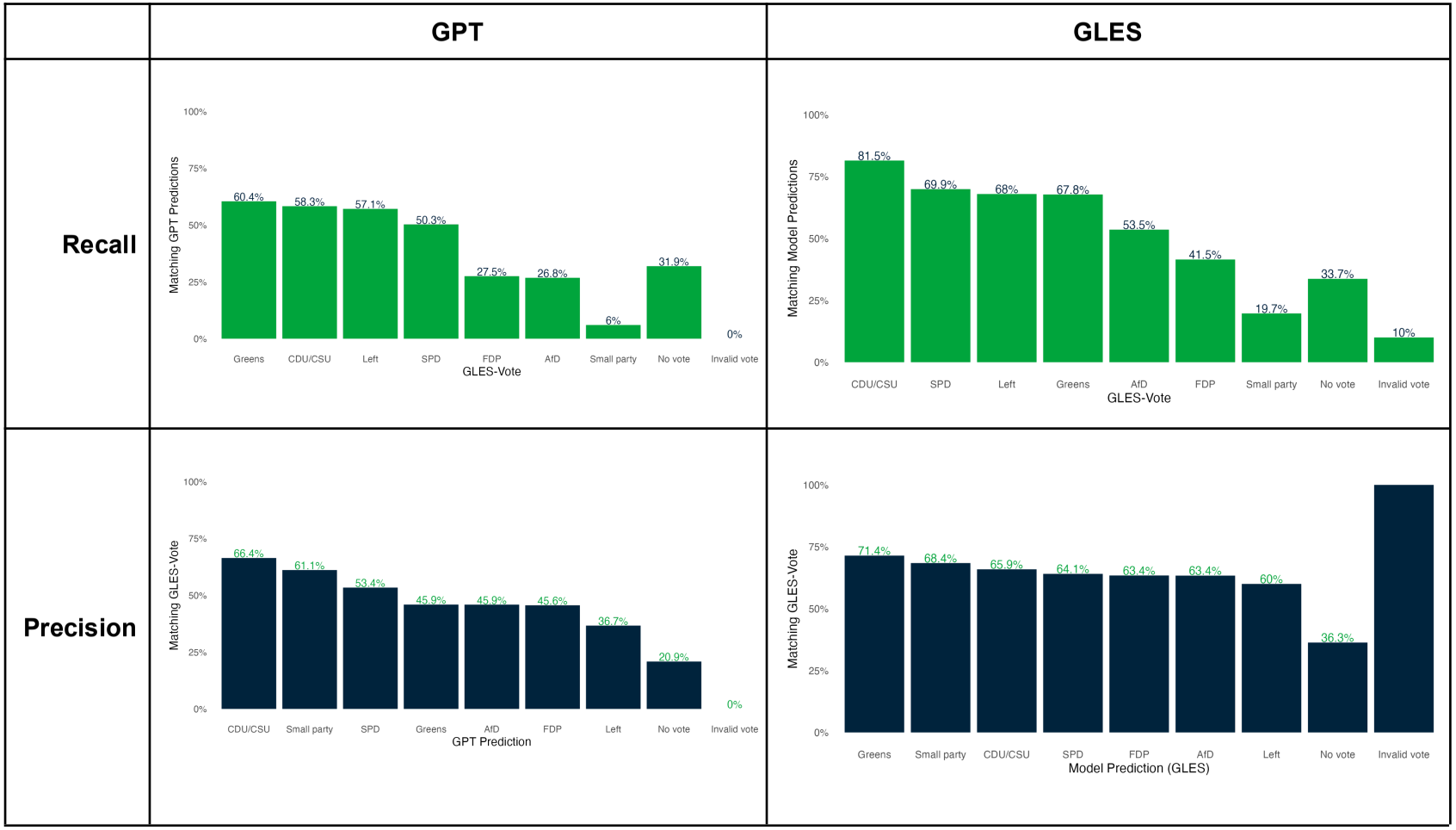


Figure A1.1: Model performance of GPT predictions vs. predictions based on multinomial regression of prompt variables on GLES-reported vote choice.

Recall refers to the number of correct predictions for a given party (true positives) divided by the number of actual votes for that party (true positives + false negatives), measuring the share of votes for a party as reported in the GLES that were correctly identified by GPT/the GLES-based model.

Precision refers to the number of correct predictions for a given party (true positives) divided by the number of all predicted votes for that party (true + false positives), measuring the share of GPT/GLES predicted votes for a given party that were correct.

F1 scores evaluate machine learning model accuracy by considering model precision and recall. The F1 score is defined as 2*precision*recall / (precision + recall), with a range of [0;1].

In imbalanced datasets – such as here, where CDU/CSU and SPD received many more (reported) votes than the other parties –, regular F1 scores might be misleading if the LLM has a tendency of assigning the modal category (i.e., the party receiving the most votes as reported by GLES), not allowing us to assess how well it performs on minority parties. Macro F1 accounts for such skews by first calculating F1 scores for each party and then averaging over all parties, thereby ensuring all parties are weighted equally regardless of their prevalence. We also explicitly include missing predictions in order not to artificially inflate the F1 scores in cases where most votes for a party were not predicted, but the remainder predicted correctly.

	F1 Score $\frac{2\times precision\times recall}{precision + recall}$	
Party	**GPT**	**GLES Model**
Overall (macro F1)	*0.39*	*0.52*
CDU/CSU	0.62	0.73
SPD	0.52	0.67
Greens	0.52	0.70
Left	0.45	0.64
FDP	0.34	0.50
AfD	0.34	0.58
No vote	0.25	0.35
Small party	0.11	0.31
Invalid	0	0.18

Table A1.6: Model accuracy (F1 scores) of GPT predictions vs. predictions based on multinomial regression of prompt variables on GLES-reported vote choice.

A1.8 Multinomial Regression Model

Values for GPT are average values across separate regressions for each of the five samples of completions.

Independent Variables		Effect on Vote Choice – GLES vs. GPT (Reference: CDU/CSU)							
		SPD	Greens	FDP	Left	AfD	Small party	Invalid vote	No vote
Age		0.000	-0.001	-0.007	-0.010	-0.003	-0.024	-0.057 *	-0.037 ***
		-0.016	-0.027	-0.014	-0.027 *	-0.019	-0.030	-0.010	-0.019
Gende (Ref.: male)	female	-0.100	-0.102	0.023	-0.475	-0.833 **	-0.239	-1.492	-0.181
		0.287	0.666 *	-0.131	0.480	-0.048	0.326	-0.674	-0.893 ***
Education		0.059	0.321 **	-0.020	0.364 **	0.156	0.106	0.253	-0.417 ***
		-0.003	0.172	0.130	-0.013	0.022	0.009	0.101	-0.397 **
Employment (Ref.: working)	studying	-0.152	0.099	-0.324	-0.445	-0.953	-0.167	0.829	-0.177
		-0.102	-0.096	-0.297	-0.057	-1.209	0.603	-0.358	1.256 *
	not working	0.399	-0.049	0.302	0.442	-0.268	-0.767	1.365	0.735 *
		-0.293	-0.209	-0.732	-0.075	-0.329	-0.267	0.076	1.560 ***
Income (Ref.: low)	medium	-0.497	-0.307	0.302	-0.539	-0.540	-0.671	-0.624	-1.006 ***
		-0.234	-0.233	-0.061	-0.441	-0.359	-0.756	-0.614	-0.824 *
	high	-0.807	-0.195	0.766	-0.843	-0.875	-1.155	-2.710	-1.037 *
		-0.406	-0.418	-0.073	-0.883	-0.353	-1.778	-0.186	-0.933
Residence (Ref.: West Germany)	East Germany	-0.331	-0.843 **	-0.194	0.263	0.269	0.275	0.874	0.276
		0.518	-0.439	-0.732	1.507 ***	1.677 ***	0.280	0.305	0.197
Religiosity (Ref.: not at all)	not very	-0.201	0.179	0.276	0.067	-0.668	0.117	-3.483	-0.447
		-0.410	-0.380	-0.461	-0.534	-0.426	-0.529	-0.512	-0.767
	somewhat	-0.244	-0.184	-0.256	-0.641	-0.829 **	-0.526	-0.224	-0.408
		-0.883 *	-1.465 ***	-1.122 *	-1.346 ***	-0.377	-0.664	-1.323	-1.436 ***
	very	-1.022 **	-0.460	-0.744 *	-1.303 *	-1.460 **	0.052	-2.900	-0.776 *
		-1.515 **	-1.617 **	-0.905	-1.553 **	-0.318	-1.542	-1.544	-1.003 *
LR-Ideology		-0.437 **	-0.554 **	0.197	-1.154 ***	0.766 ***	-0.462	-0.631	0.162
		-0.787 **	-1.331 ***	-0.269	-1.938 ***	0.919 **	-0.680	-0.287	-0.677 **

cont.

Independent Variables		Effect on Vote Choice – GLES vs. GPT (Reference: CDU/CSU)							
		SPD	Greens	FDP	Left	AfD	Small party	Invalid vote	No vote
Party ID (Ref.: CDU/CSU)	SPD	1.064	1.885	3.178	3.384	1.828	-1.377	3.489	0.205
		-7.848	-7.043	-3.839	-7.208	-6.585	-2.030	-1.673	-3.380
	Greens	0.003	4.195	5.674	-1.611	-1.826 **	2.201	-1.375	-9.351 *
		0.653	-3.556	2.219	-2.114	-2.775	1.455	0.023	0.839
	FDP	-0.317	0.721	5.394 **	1.896	-0.017	-5.423	0.132	3.549
		-4.024	-4.153	-2.779	-0.526	-5.213	-0.189	-0.603	0.596
	Left	-2.236	7.655	4.615	1.456	1.244	-0.259	1.161	-0.049
		-0.344	-2.915	-1.054	-3.084	0.683	1.790	0.201	2.835
	AfD	6.891	-1.832 *	-1.548 **	-1.154	4.123	-2.205 *	-2.487	2.558
		-1.040	-4.932	-1.126	-5.672	-9.735 **	-0.501	-0.679	-2.304
	Small party	-0.877 ***	8.022	-0.184 ***	0.427	-1.809	-1.845	4.214	-0.741
		3.897	-5.410	0.572	-3.173	-2.134	-0.349	0.335	1.884
	none	-0.588	1.215	1.037	1.320	-0.580	-1.940	-0.175	-1.527
		-2.881	-5.475 **	-1.956	-5.762 ***	-4.704 **	0.895	1.056	-0.739
Strength of Party ID		-0.954 *	-0.309	-0.179	-1.167 *	-0.758 *	-1.417 **	-1.405	-1.020 ***
		-1.767 **	-2.564 ***	-1.227	-2.658 ***	-1.963 ***	-0.504	-0.762	-1.439 **
Party ID x Strength (Ref.: CDU/CSU)	SPD	1.241 *	0.351	-0.509	0.639	0.033	1.238	0.480	0.630
		4.291 **	3.435 **	1.829	3.503 *	2.832	1.911	1.422	2.007
	Greens	0.936	0.308	-1.448	1.893 *	-0.410	0.237	0.905	2.811 *
		0.898	3.018 **	0.207	1.263	0.944	0.810	0.899	0.847
	FDP	0.561	0.384	-0.423	0.795	0.353	2.251	0.444	-1.155
		1.707	2.280	2.769	0.112	2.050	1.071	0.973	0.952
	Left	1.710	-1.347	-0.916	1.972 *	0.602	1.412	1.230	1.013
		1.980	2.729	0.274	3.488	0.950	1.467	1.592	1.154
	AfD	-1.923	-0.013	-0.191	0.427	0.369	0.190	2.511	0.271
		0.634	1.063	0.763	2.182	4.016 ***	0.759	0.437	1.421
	Small party	-0.533 ***	1.368	0.982 ***	3.328	4.656 **	6.076 ***	4.457 **	4.779 ***
		-0.963	4.514 **	1.124	3.192 *	3.456 *	4.610 **	1.499	2.532

cont.

Independent Variables		Effect on Vote Choice – GLES vs. GPT (Reference: CDU/CSU)							
		SPD	Greens	FDP	Left	AfD	Small party	Invalid vote	No vote
Att. Inequality (Ref.: take measures)	no opinion	-0.350	-1.076 **	-0.427	-0.984 *	-0.519	-0.226	-3.204	-0.500
		-0.692	-0.730	0.051	-0.315	0.182	-0.623	-0.411	0.739 *
	don't take measures	-0.888 *	-0.692	0.225	-0.810	-0.145	-0.651	-0.111	-0.503
		-1.893 ***	-2.228 ***	-0.317	-2.049 **	-0.105	-2.664	-1.973	-0.065
Att. Immigration (Ref.: facilitate)	neither nor	-0.572	-0.843 *	0.173	-0.456	0.438	-0.286	1.856	0.088
		-0.387	-0.432	-0.024	-0.253	-1.423	-0.035	-0.633	-0.862
	limit	-0.513 *	-0.987 ***	0.045	0.000	1.435 ***	-0.219	2.014	0.211
		-0.890 **	-1.220 ***	-0.238	-0.712 *	-0.155	-0.176	-0.823	-0.739 *
Intercept		1.224	-1.689	-1.175	-1.671	0.250	2.753	-1.528	5.117 ***
		5.392 *	8.490 ***	3.080	8.353 ***	4.816	-3.755	-2.624	6.081 ***

N	1827
*	p<0.05
**	p<0.01
***	p<0.001

Multinomial regression.
Reported values represent beta coefficients.

Table A1.7: Results of a multinomial regression of prompt variables on vote choice.

A1.9 Distribution of partisanship and choice in regression samples

FDP and AfD are the parties with the lowest shares of partisans in the survey data, with the reported vote share for these parties being twice as high as their respective share of partisans. Moreover, the share of non-voters is disproportionately high among people identifying with the AfD or a small party, and non-partisans.

Party ID **Vote**	**CDU/ CSU** (28.9%)	**SPD** (19%)	**Greens** (10.1%)	**FDP** (5.3%)	**Left** (8.6%)	**AfD** (4.5%)	**Small party** (1.1%)	**No party** (22.4%)
CDU/CSU (26.9%)	**70.8%**	6%	5.4%	11.5%	1.9%	2.4%	0%	17.1%
SPD (18%)	2.8%	**67%**	8.1%	2.1%	6.3%	1.2%	0%	13%
Greens (11.7%)	2.3%	6%	**75.1%**	3.1%	5.1%	**0%**	5%	7.3%
FDP (10.3%)	9.7%	3.2%	1.6%	**74%**	1.3%	0%	0%	12.2%
Left (9.9%)	0.2%	4.9%	4.3%	2.1%	**67.7%**	0%	0%	11.2%
AfD (8.5%)	5.3%	2.6%	**0%**	3.1%	3.2%	**77.1%**	5%	11%
Small party (3.6%)	1.1%	1.7%	3.2%	2.1%	5.7%	0%	**65%**	5.9%
Invalid (0.5%)	0%	0.6%	0%	0%	0.6%	1.2%	**5%**	1.2%
No vote (10.6%)	7.8%	8%	**2.2%**	**2.1%**	8.2%	**18.1%**	**20%**	**21%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Table A1.8: Distribution of partisanship and reported vote choice according to GLES (N=1827).
Note: Not including missing completions; Column percentages; overall totals of variables in parentheses.

A1.10 Robustness checks

1. Comparison of results using imputed and non-imputed personas

For our main analyses, we imputed missing values on any of the prompt variables for 377 respondents to obtain complete personas for prompting GPT. Most of these missing values concerned household income (248) and political ideology (114), with missing values for all other variables affecting less than 30 individuals, respectively, and no missing values for the key demographic indicators age and gender. As a robustness check, we re-prompt the LLM for these respondents using an adjusted prompt containing only the non-imputed, incomplete information (see Table A2b in Appendix II for an example prompt). We then merge these non-imputed observations and their GPT predictions with the a-priori complete cases to once again create a full sample and compare it to the main sample.

Relying on the non-imputed personas results in a higher overall share of missing predictions. While this share was 0.9% for the main sample after up to three trials, it is 2.3% for the total sample including the non-imputed personas. However, this difference has no impact on the substantive results. The aggregate vote shares estimated by GPT differ from the main results only by tenths of a percentage point, the only exception being the estimated share of non-voters now being two percentage points lower, 14.6 compared to 16.5%. The overall share and variance in matching vote choices between GPT predictions and GLES reports is the same. Using the non-imputed personas, both the matches by GLES party vote (recall) and the LLM's precision (matches by GPT prediction) per party are of similar magnitude, with minor changes in the rank order, as are the overall F1 scores per party.

As it lies in the nature of the non-imputed observations that they carry missing values on at least one predictor variable, we cannot compare the differential impact of the (missing) predictor variables on vote choice. However, we can employ a binary indicator for individuals with missing (sample with non-imputed personas) or imputed (main sample) values. Comparing multinomial regressions on GPT-predicted vote choice confirms the aggregate-level observation that for non-imputed personas, i.e., those with missing values, GPT is more likely to predict "Did not vote", an effect that is not present for the same people in the main sample including the imputed values. Considering that prompting GPT with the non-imputed personas led to a higher share of missing predictions to begin with, the effect of non-imputation on the completed predictions may even be underestimated. Distinguishing between the LLM's prediction of (individual) turnout and substantive vote choice could provide further insights into this matter. For an analysis of the effect of using imputed versus incomplete prompts on the successful completion of a prediction, see the next subchapter.

Taken together, these results suggest that GPT not only appears to be more likely to return a prediction, but also is more likely to predict a person voted when provided with more information about the person.

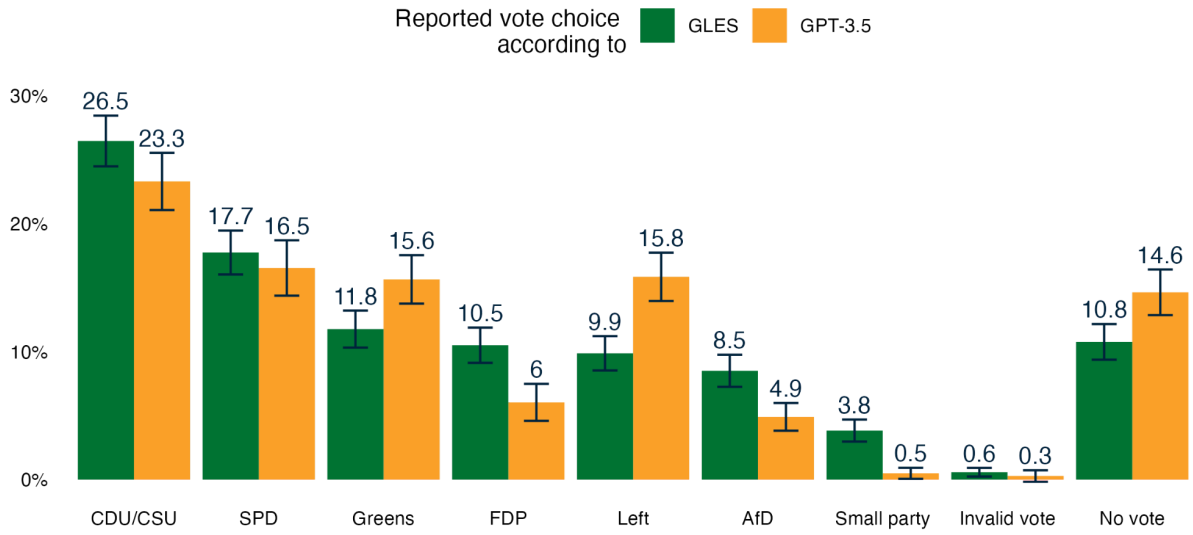


Figure A1.2: Distribution of vote shares as estimated by GLES and GPT – robustness sample.

Party	**F1 Score** $\frac{2 \times precision \times recall}{precision + recall}$
Overall (macro F1)	0.39
CDU/CSU	0.63
Greens	0.53
SPD	0.52
Left	0.45
FDP	0.35
AfD	0.34
No vote	0.23
Small party	0.12
Invalid	0

Table A1.9: Model accuracy (F1 scores) – robustness sample.

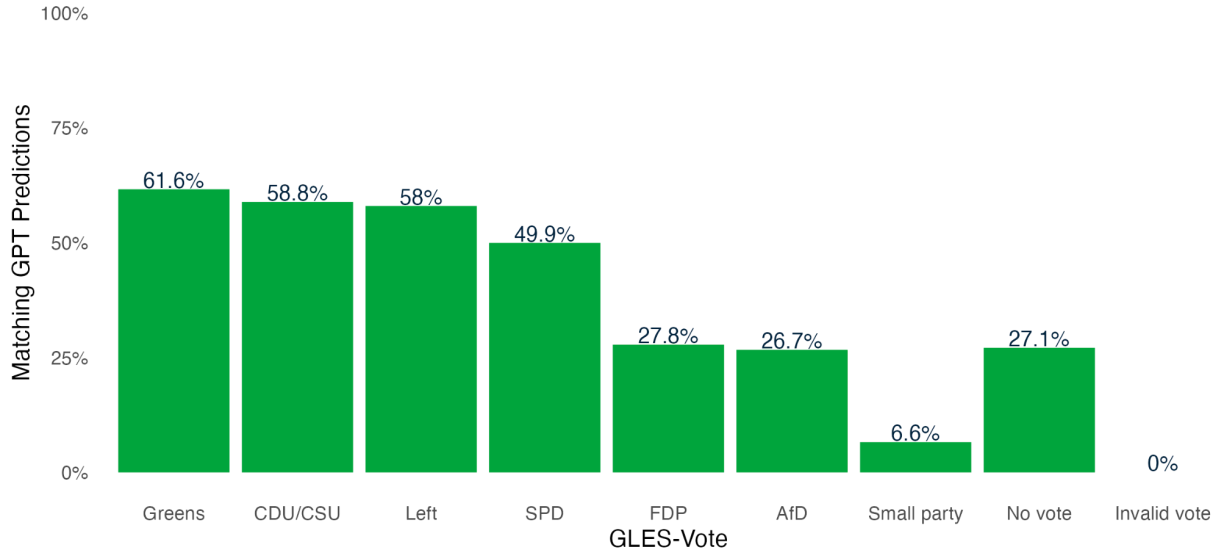


Figure A1.3a: Share of party votes according to GLES that GPT correctly predicted (recall) – robustness sample.

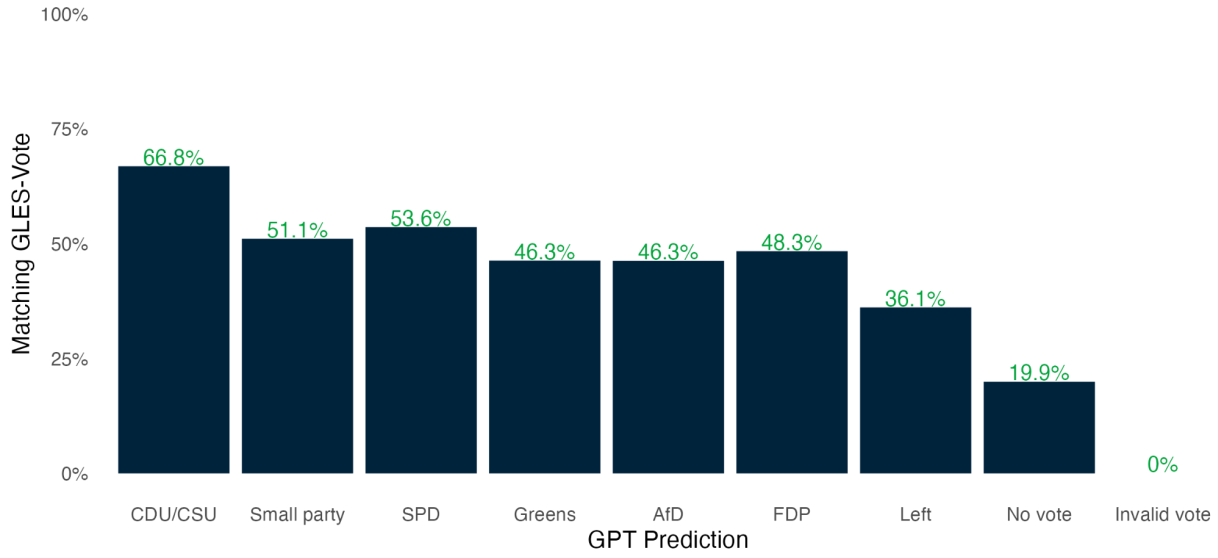


Figure A1.3b: Share of GPT predictions that voted for the respective party (precision) – robustness sample.

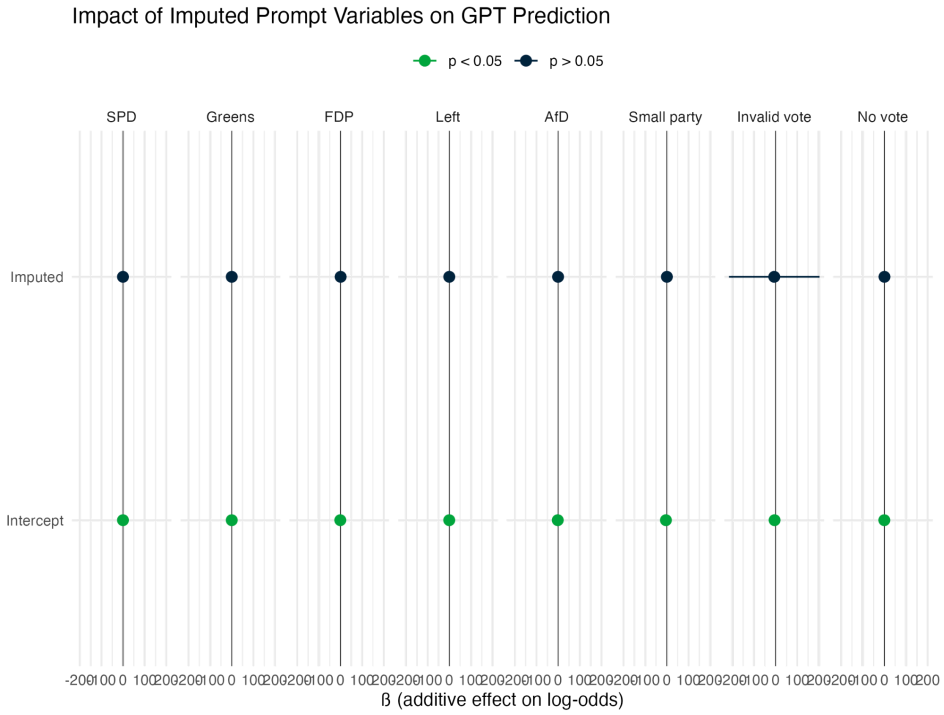


Figure A1.4: Results of a logistic regression of indicator for imputed cases on GPT vote choice – main sample.

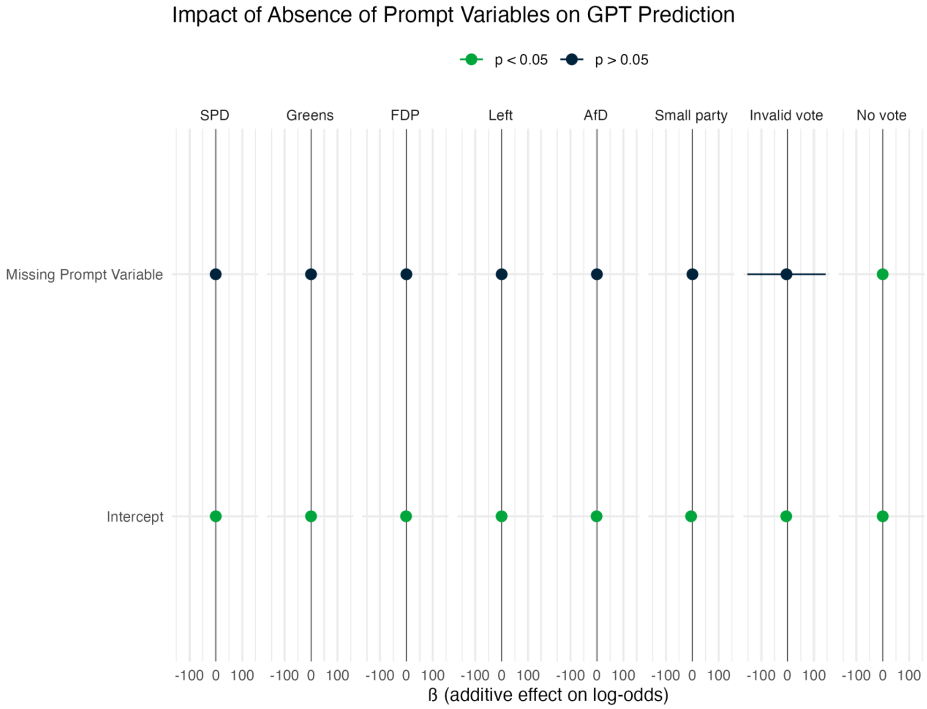


Figure A1.5: Results of a logistic regression of indicator for cases with missing prompt variables on GPT vote choice – robustness sample.

2. Analyses of missing completions

After re-sampling completions not containing an identifiable vote choice up to two times, 87 or 0.9% of completions remained without such information, corresponding to 78 individuals (4.3% of the sample), 9 of which have two missing predictions. On average, there are 15 missing predictions in each of the five iterations we ran for each persona, the fourth iteration being an outlier with 27 missing cases. For ensuring comparability across iterations, these individuals were excluded from the regression analyses. In the robustness sample, 222 completions (2.3%) do not contain a prediction, corresponding to 150 individuals (7.9% of the total sample), 114 of which have at least two and up to five missing predictions. Considering only the non-imputed personas in the robustness sample, 141 completions (7.5% of non-imputed personas) do not contain a prediction. At 43, the average number of missing predictions in each of the five iterations is much higher than in the main sample, and once again, the fourth iteration is an outlier with 50 missing cases. The differences in missing predictions between imputed and non-imputed personas support the notion that the full set of information is beneficial to a successful prediction – although the main results show that not all information is weighted equally in the outcome of the prediction.

In both the main and especially in the robustness sample, the shares of missing predictions increased over the course of the three rounds of sampling we performed for those completions that did not contain a vote choice in the previous round (see Table A4 in Appendix V). A descriptive analysis reveals that across samples, individuals for whom at least one GPT prediction was missing are on average younger than those with five predictions (47 vs. 52 years). The (binary) gender distribution among them is more balanced, while complete cases skew male. Those for whom a prediction was missing tend to be employed more often, resulting in fewer non-working individuals. The share of low-income individuals is lower among those with missing predictions, while that of medium-income individuals is higher. Ideologically, individuals with missing predictions tend to position themselves more in the middle, while there are more strongly-left and -right-leaning individuals among complete cases. Regarding partisanship, those with missing predictions largely do not identify with any party at all, much more so than complete cases. Notably, there are no Green or AfD partisans, but more small party partisans. Moreover, among those who do identify with a party, many more incomplete cases do so only weakly, compared to complete cases. The share of those supporting immigration is much lower among those with missing predictions. While these patterns are similar across the main and robustness samples, there are some differences regarding other prompting variables. For example, while the share of East German residents is higher among individuals with missing predictions in the main sample, the opposite is true in the sample containing the non-imputed cases. Although in both samples, incomplete cases tended to vote less for the CDU/CSU and SPD and more for the FDP, AfD (according to their survey response), only in the robustness sample was the share of those voting for the Greens lower and that of those voting for the AfD, small parties, or not voting at all, much higher among individuals with missing predictions. Interestingly, in the main sample (containing imputations), individuals with missing predictions are less likely to be imputed cases.

In sum, GPT appears to be more likely to make complete predictions for older, male, wealthier individuals who are ideologically unambiguous, strong (especially Green or AfD) partisans or voted for one of the bigger, centrist parties, and tend to support immigration. This echoes the bias observed in our main analyses, indicating that GPT tends to pick up on signals representing dominant or highly "visible" subgroups, while struggling with non-typical subgroups.

		Main sample (including imputed values)		**Robustness sample (no imputed values)**	
Subgroup		Individuals with 5 completions (N = 1827)	Individuals with at least one missing completion (N = 78)	Individuals with 5 completions (N = 1755)	Individuals with at least one missing completion (N = 150)
Age	average	47	51	47	52
Gender	male	52.7%	50%	52.7%	50.7%
	female	47.3%	50%	47.3%	49.3%
Education	no degree	1.5%	1.3%	1.5%	2%
	Hauptschule	21.5%	17.9%	21.4%	20.8%
	Realschule	31.2%	30.8%	31.2%	30.9%
	Abitur	17.2%	19.2%	17.2%	19.5%
	College	28.5%	30.8%	28.8%	26.8%
Employment Status	not working	35.7%	23.1%	36.1%	26%
	studying	7.5%	7.7%	7.4%	8.7%
	working	56.8%	69.2%	56.5%	65.3%
Income	low	18.7%	9%	17.5%	15.8%
	medium	69.1%	78.2%	69.8%	72.6%
	high	12.2%	12.8%	12.7%	11.6%
Residence	West	67.9%	62.8%	67.4%	70.7%
	East	32.2%	37.2%	32.6%	29.3%

Religiosity	not at all	35.1%	33.3%	35.7%	29.5%
	not very	16.6%	20.5%	16.4%	21.9%
	somewhat	36%	39.7%	35.4%	43.2%
	very	12.2%	6.4%	12.4%	5.5%
cont. **Ideology**	strongly left	8.2%	3.8%	8.6%	2.3%
	center-left	26.1%	24.4%	26.7%	20.8%
	in the middle	55%	66.7%	53.5%	70.8%
	center-right	8.8%	5.1%	9.5%	5.4%
	strongly right	2%	0%	2%	0.8%
Party ID	CDU/CSU	28.9%	15.4%	30%	9.5%
	SPD	19%	10.3%	19.5%	6.6%
	Greens	10%	0%	10.6%	0%
	FDP	5.3%	2.6%	5.3%	1.5%
	Left	8.6%	1.3%	9%	0.7%
	AfD	4.5%	0%	4.7%	0%
	Small party	1.1%	3.8%	1.1%	2.9%
	none	22.4%	66.7%	19.8%	78.8%
Strength of Party ID	none	22.4%	66.7%	20.4%	81.2%
	very weak	0.6%	66.7%	0.6%	0%
	rather weak	2.5%	0%	2.5%	2%
	moderate	30.4%	2.6%	30.9%	7.4%
	rather strong	37.7%	12.8%	39%	8.1%
	very strong	6.4%	16.7%	6.6%	1.3%

Att. Inequality	don't act	10.2%	3.8%	10.1%	7.6%
	no opinion	13%	17.9%	13.1%	14.5%
	act	76.8%	78.2%	76.7%	77.9%
Att. Immigration	limit	49.8%	53.8%	49.6%	52.4%
	neither nor	18.2%	23.1%	18.3%	21%
	facilitate	32%	23.1%	32.1%	26.6%
Vote choice (GLES)	CDU/CSU	26.9%	16.7%	27.2%	17.3%
	SPD	18%	11.5%	18.1%	13.3%
	Greens	11.7%	12.8%	12%	9.3%
	FDP	10.3%	15.4%	10.4%	11.3%
	Left	9.9%	9%	9.7%	11.3%
	AfD	8.5%	9%	8.4%	10%
	Small party	3.6%	9%	3.6%	6.7%
	Invalid	0.5%	1.3%	0.6%	6.7%
	No vote	10.6%	15.4%	10%	20%
Imputed	No	80%	92.3%	100%	100%
	Yes	20%	7.7%	0%	0%

Table A1.10: Distributions of subgroup characteristics by missing completion and imputation status (column percentages per variable; rounded values).

3. Analyses without misclassified respondents

In some instances, GPT misclassified respondents as non-voters because it considered them ineligible to vote, mostly based on their age (all German citizens aged 18 and over are eligible to vote).

In 16 completions (14 unique respondents, 0.7% of the sample), the LLM wrongly stated that the respective respondent (between 58 and 94 years old) was too old to be eligible to vote and therefore did not vote.

Moreover, for 51 completions (44 unique respondents, 2.3% of the sample), GPT assumed respondents to be too young and therefore ineligible to vote. Only half of these cases (24, corresponding to 19 unique respondents) can be attributed to the fact that we did not specify that the prompt information referred to 2017, thereby possibly inducing GPT to adjust for the time difference of six years between the election and prompting. However, in the other half of cases (27 / 25 respondents), the respondent would have been over 18 in 2017 even if assuming the age information in the prompt referred to 2023, and in 16 of those cases, the respondent was between 30 and 94 years old.

Finally, GPT considered 17 cases (and respondents, 0.9% of the sample) ineligible to vote and therefore predicted "did not vote" without giving a specific reason for ineligibility.

However, all of these instances occurred at most twice per respondent, with the remaining three or four GPT completions per respondent containing a different vote choice. In total, 18 of the 73[1] respondents (25%) who GPT wrongly considered ineligible to vote and predicted to be non-voters because of this actually did not vote according to the GLES data.

Excluding these respondents from the analysis (along with those with less than five complete predictions) yields the following results: Even though the estimated vote shares change in absolute values for both the GLES and GPT data due to the omitted respondents, the relative differences remain the same. There continue to be no significant differences for the CDU/CSU and SPD. Even when excluding some of the individuals GPT had predicted to be non-voters, GPT significantly overestimates the share of Green, Left, and non-voters, and underestimates those of FDP, AfD, and small party voters. The F1 scores change slightly, but only on the second decimal, and do not substantially change the ranking.

[1] Two respondents were considered generally ineligible in one completion, too old in the other, hence being counted towards both groups but only once in the overall total.

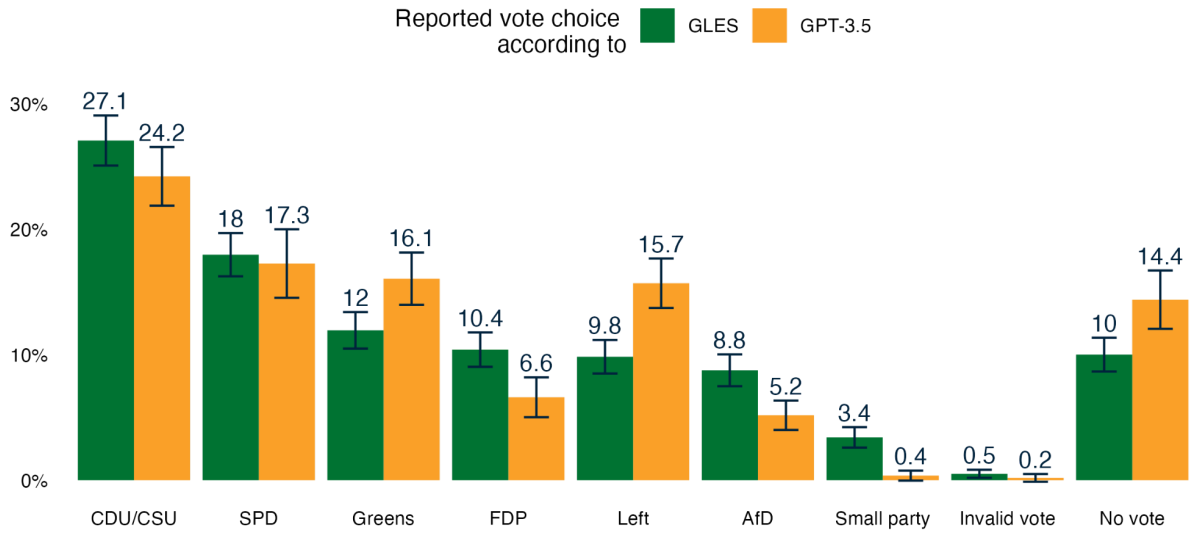


Figure A1.6: Distribution of vote shares as estimated by GLES and GPT, excluding respondents considered ineligible.

Party	**F1 Score** $\frac{2\times precision\times recall}{precision+recall}$
CDU/CSU	0.63 (+0.01)
Greens	0.54 (+0.02)
SPD	0.53 (+0.01)
Left	0.46 (+0.01)
FDP	0.36 (+ 0.02)
AfD	0.35 (+ 0.02)
No vote	0.22 (- 0.03)
Small party	0.12 (+ 0.01)
Invalid	0

Table A1.11: Model accuracy (F1 scores), excluding respondents considered ineligible.

A2 Appendix to Chapter 3

A2.1 Prompting (variables, translations, parties, keywords; election results)

see electronic dissertation file or here: https://github.com/leahvdh/dissertation_lvdh/

A2.2 Summary statistics per country (prompt variables and number of missing values)

see electronic dissertation file or here: https://github.com/leahvdh/dissertation_lvdh/

A2.3 Country selection

Country	**Language Internet Coverage [source]**	**Language Speakers in EU [source]**	**Language Family [source]**	**Population Size (millions) [source]**	**European Region [source]**	**Political Position within Europe (Shapley-Shubic Index of Bargaining Power) [source]**	**Share of support for EU member-ship [EB 99.4: QA12.2]**
France	4.4%	25%	Romance	68.2	West	13.68	55.3%
Germany	5.4%	29%	Germanic	83.1	West	17.23	68.9%
Poland	1.8%	9%	Slavic	36.8	East	7.08	50.4%
Slovakia	0.4%	2%	Slavic	5.4	East	1.61	62.3%
Sweden	0.5%	3%	Germanic	10.5	North	2.39	78.5%

Table A2.1: *Variation in selected countries' linguistic, geographic, political, and attitudinal contexts.*

A2.4 Additional Analyses: Turnout

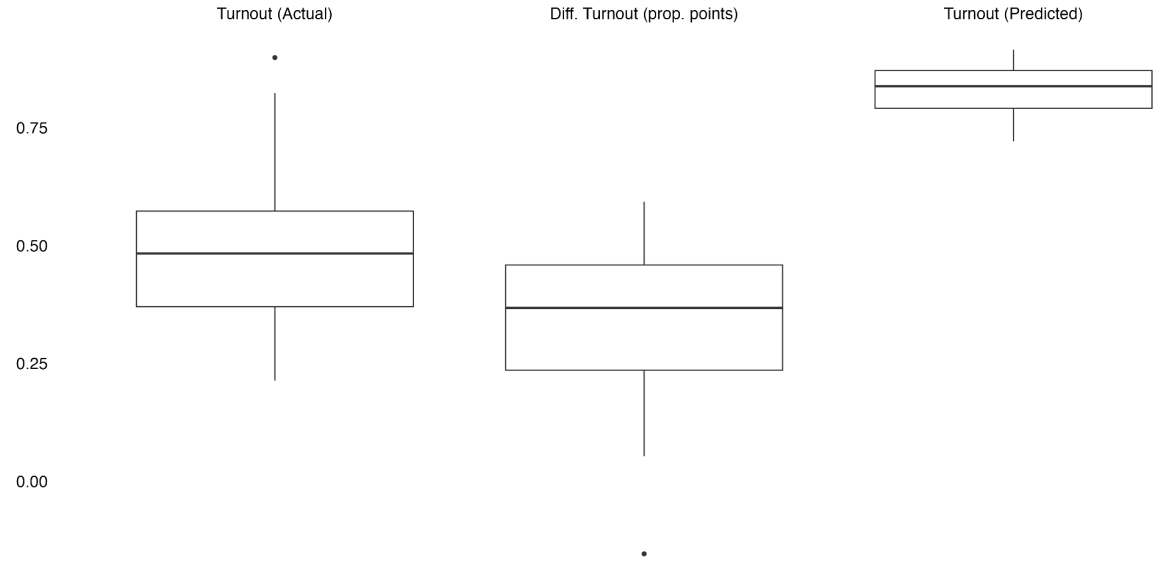


Figure A2.1: *Distribution of actual and predicted turnout and their differences (based on full English prompt).*

A2.5 Additional Analyses: Party Vote Shares

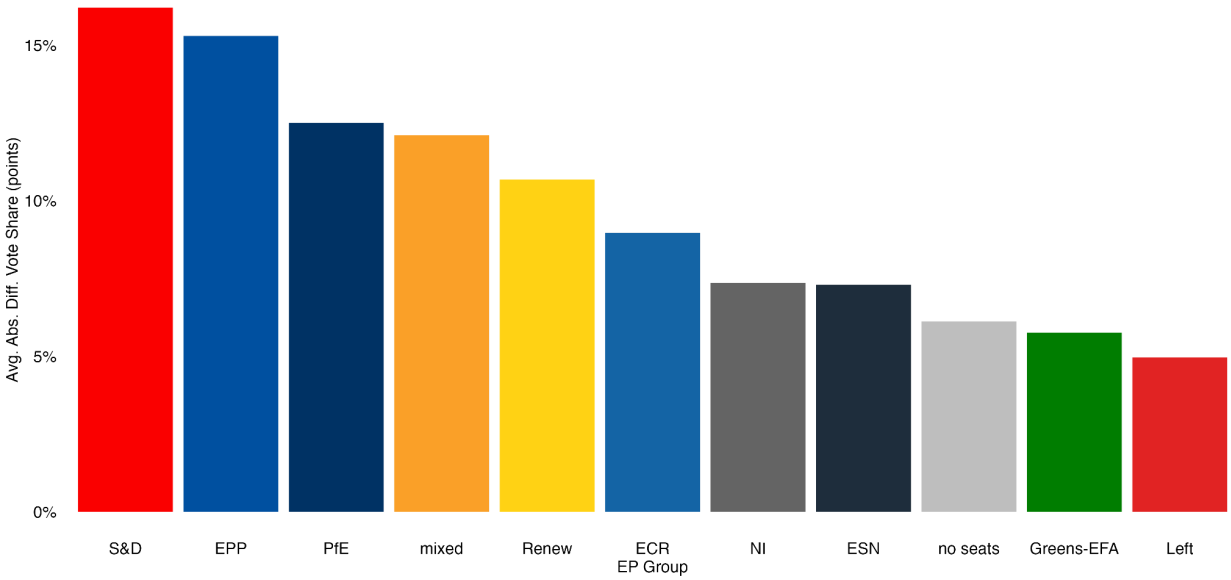


Figure A2.2: *Average absolute differences in party vote shares by EP group (based on full English prompt).*

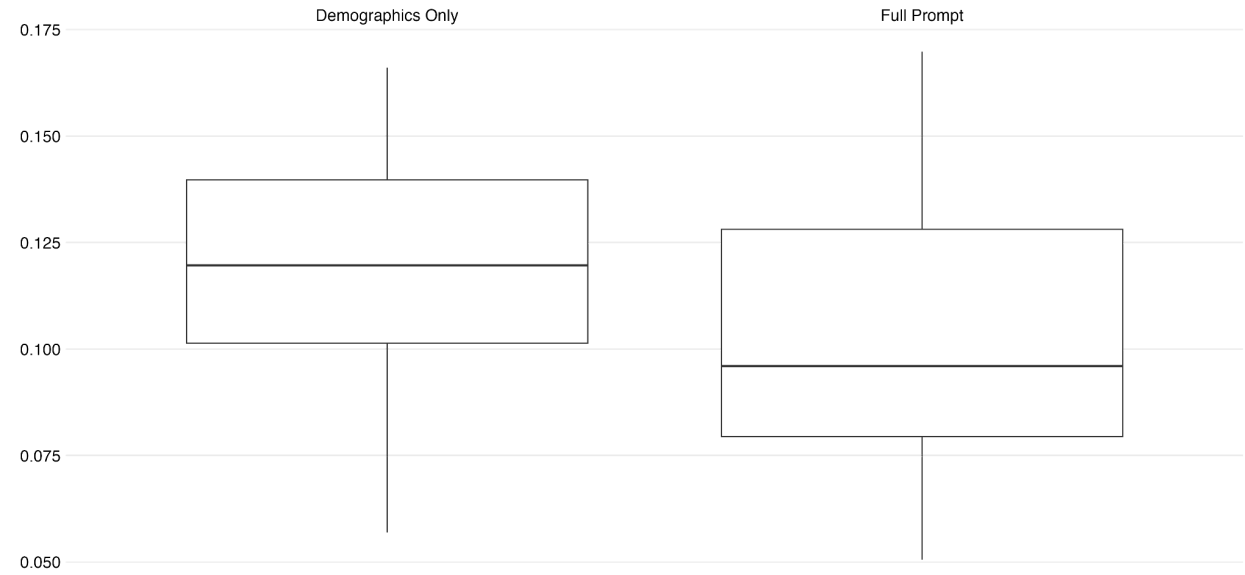


Figure A2.3: *Distribution of per-country average absolute differences in party vote shares by prompt content, as proportion points (based on English prompt).*

A2.6 Analyses for Open-Source Models

1. Llama 3.1

Note: Compared to GPT, the cross-country average share of missing data is 114 times higher for predictions of turnout (0.02% vs. 1.8%) and 15 times higher for predictions of party vote shares (0.2% vs. 2.5%). This is the case especially for Poland and Slovakia as well as the other countries whose average includes the completions based on native-language prompts which contained a higher amount of missing values.

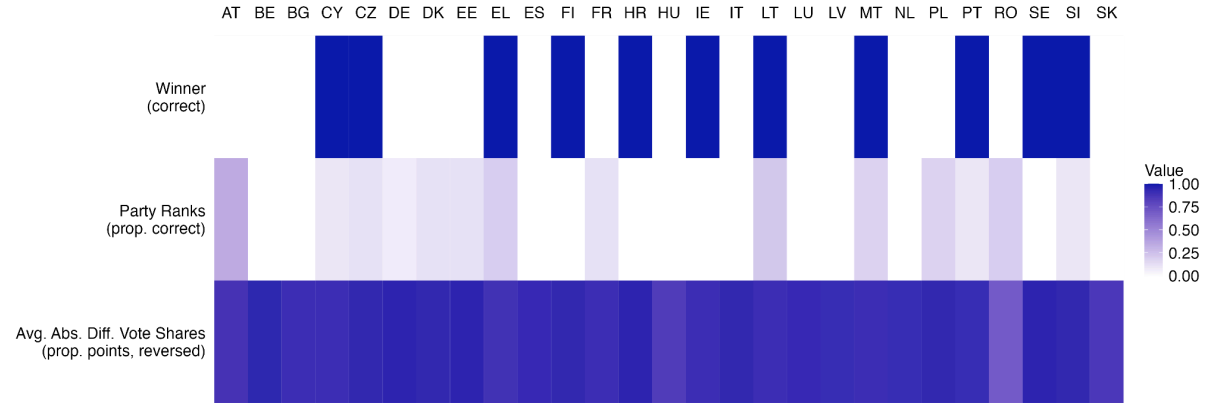


Figure A2.4: *Predictive performance of Llama 3.1 for the 2024 EU election party results (based on full English prompt).*

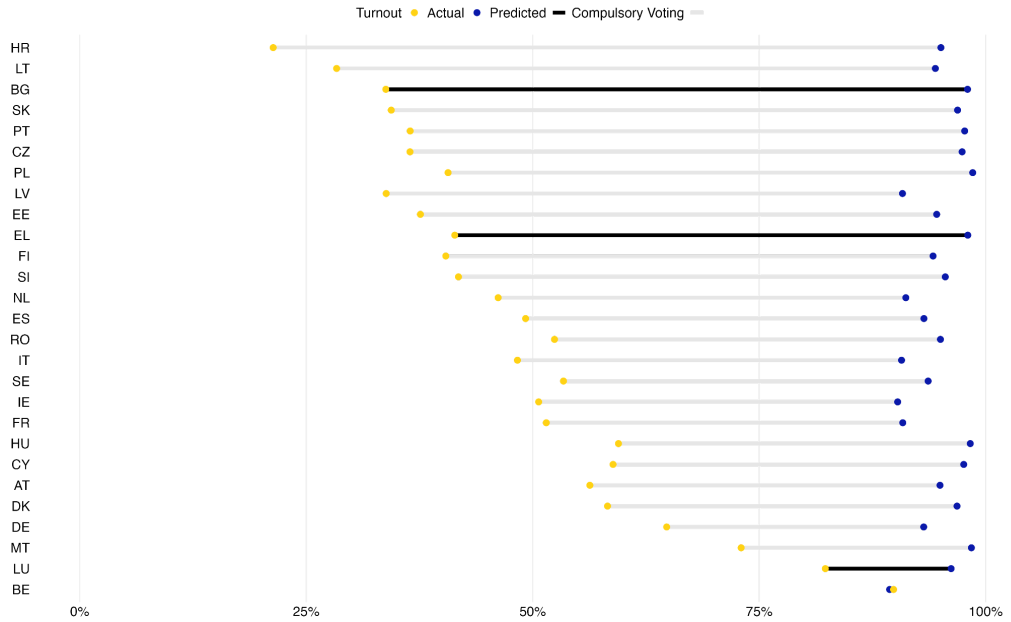


Figure A2.5: *Difference between actual turnout in the 2024 EU elections and Llama 3.1's predictions (based on full English prompt).*

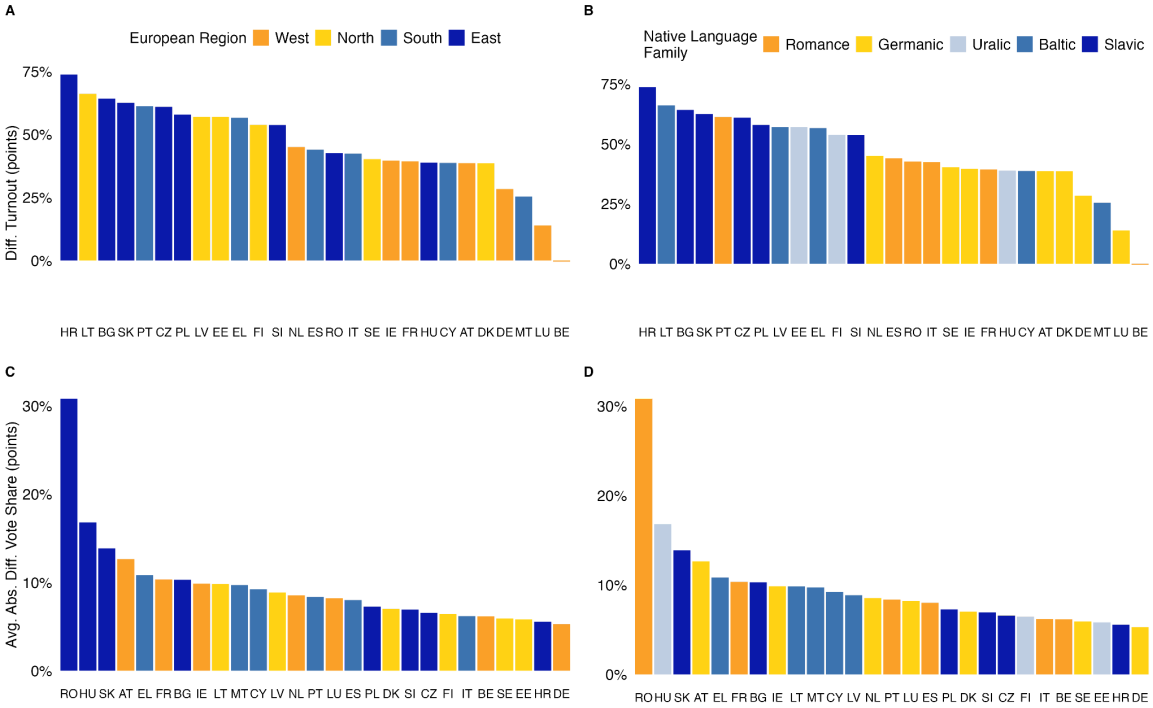


Figure A2.6: *(Average) difference between actual turnout and party vote shares in the 2024 EU elections and Llama 3.1's predictions (based on full English prompt)* ***by region and language family***.

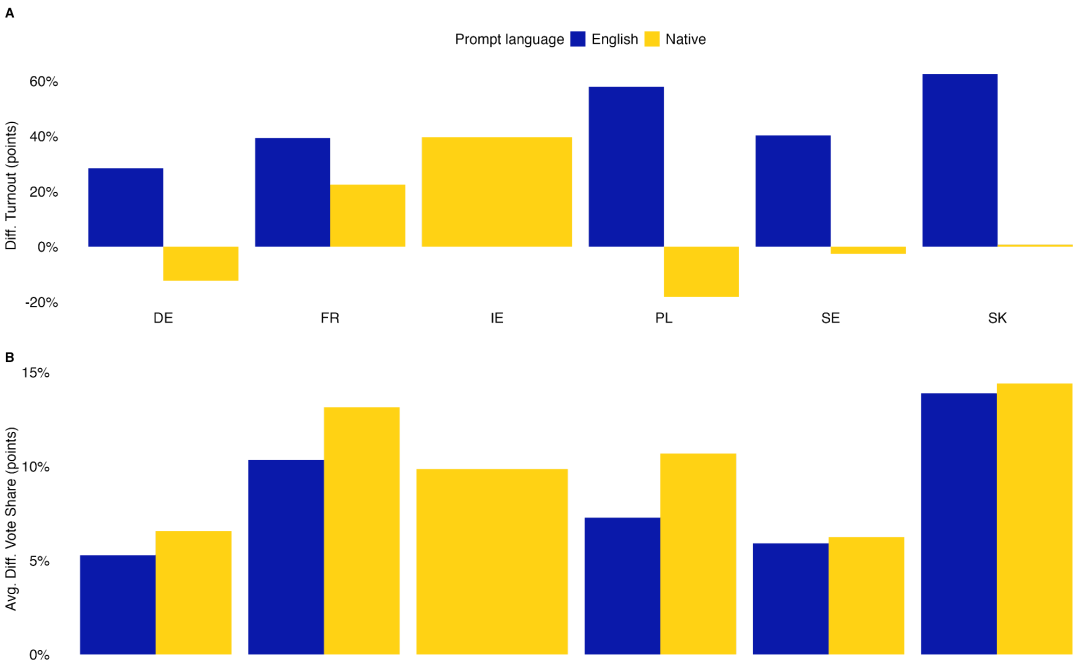


Figure A2.7: *(Average) difference between actual turnout and party vote shares in the 2024 EU elections and Llama 3.1's predictions (based on full prompt)* ***by prompt language***.

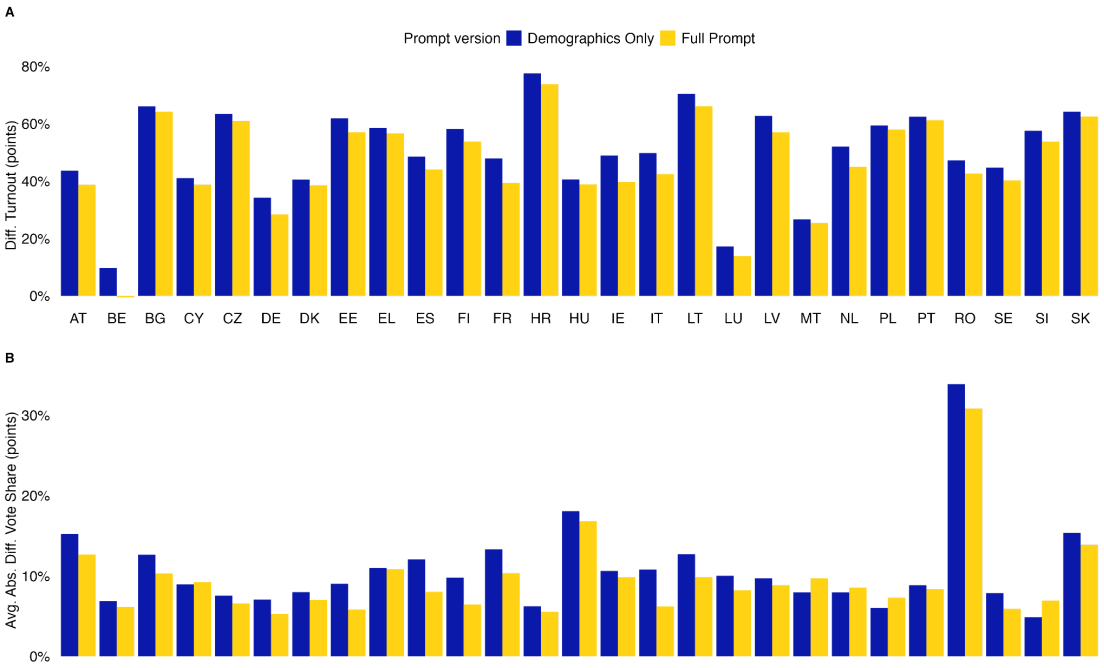


Figure A2.8: *(Average) difference between actual turnout and party vote shares and predictions using Llama 3.1 (based on **English** prompt) **by prompt content**.*

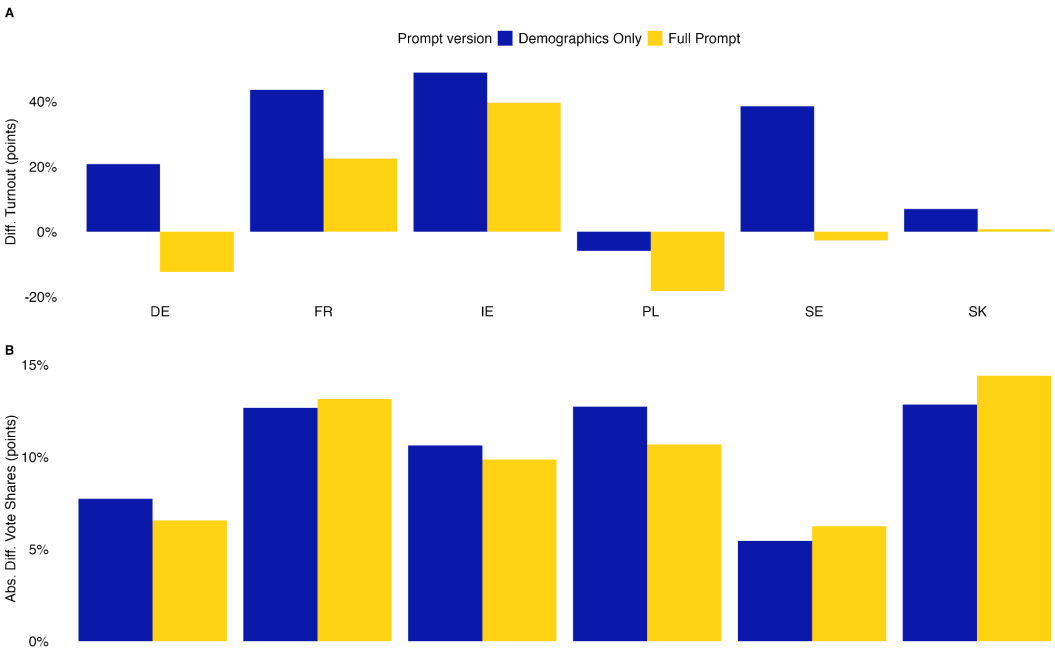


Figure A2.29: *(Average) difference between actual turnout and party vote shares and predictions using Llama 3.1 (based on **native language** prompt) **by prompt content**.*

2. Mistral

Manual checks of the automated extractions of vote choices revealed that a disproportionate amount of completions lacked a vote choice. For 11 out of 64 country-language-prompt version-permutations, systematic manual checks of the automated vote choice extraction confirmed the large amount of missing values, upon which it was decided not to pursue further analyses. The sample of manual checks included English and native-language prompting, full-information and demographics-only prompting, Eastern and Western European countries, countries with Slavic, Romance, and Germanic native languages, and countries using Arabic and Cyrillic alphabets, corroborating that this is a general issue with Mistral. Notably, there are fewer missing values in the datasets using German prompting, and more missing values in those using a demographics-only prompt.

Country	**Language**	**Prompt Version**	**Share NAs: Turnout**	**Share NAs: Party Choice (of non-NAs for turnout)**
AT	EN	full	46.2%	34.4%
AT	EN	dem.	71.8%	8.8%
BE	EN	full	50.2%	68.3%
BE	EN	dem.	75.2%	66.4%
BG	EN	full	41.8%	46.2%
BG	EN	dem.	72.9%	18.2%
CY	EN	full	55.0%	77.4%
CY	EN	dem.	69.5%	46%
CZ	EN	dem.	77.8%	45.6%
DE	DE	full	0.9%	14.3%
DE	DE	dem.	3.0%	16.8%

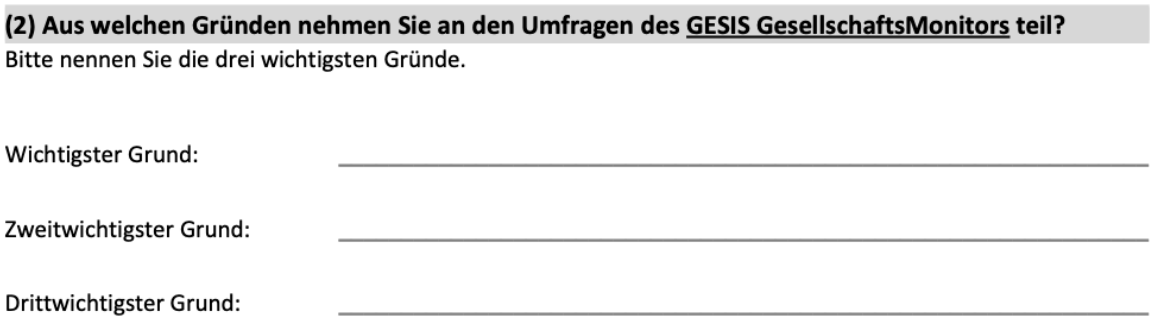
Table A2.2: *Proportions of missing values in selected datasets using Mistral.*

A3 Appendix to Chapter 4

A3.1 Coding scheme, descriptions and examples in German and English

see electronic dissertation file or here: https://github.com/leahvdh/dissertation_lvdh/

A3.2 *Survey Question*


(2) Aus welchen Gründen nehmen Sie an den Umfragen des GESIS GesellschaftsMonitors teil?
Bitte nennen Sie die drei wichtigsten Gründe.

Wichtigster Grund: _____

Zweitwichtigster Grund: _____

Drittwichtigster Grund: _____

Figure A3.1: Question on survey motivation as asked in the GESIS Panel.pop (Bosnjak et al., 2018, GESIS, 2024).
Translation: "(2) For what reasons do you participate in the surveys of the GESIS GesellschaftsMonitor? Please name the three most important reasons. Most important reason: … Second most important reason: … Third most important reason: … "

A3.3 Prompting

Du bist eine Expertin für Umfragen, die offene Antworten auf die Frage, wieso Personen an einer Umfrage teilnehmen, klassifiziert. Ordne diese Teilnahmegründe genau einer der folgenden Kategorien zu.

Die Kategorien sind:

[GERMAN CATEGORY 1: German description]
[GERMAN CATEGORY 2: German description]
[...]
[GERMAN CATEGORY 22: German description]

Stelle deine bestmögliche Vermutung an, auch wenn es schwer fällt.
Antworte im folgenden Format: Teilnahmegrund | KATEGORIE.
Begründe deine Zuordnung nicht, sondern gib nur den Teilnahmegrund und deine Zuordnung zurück.

Beispiele:
[German example reason | GERMAN CATEGORY 1]
[German example reason | GERMAN CATEGORY 2]
[...]
[German example reason | GERMAN CATEGORY 22]

Klassifiziere den folgenden Teilnahmegrund:

[German open-ended response]

Figure A3.2: German prompt used for LLM-based classifications of the open-ended survey question. Categories and, in the detailed approach, descriptions (green font) were randomized across individual queries. In the few-shot approach, examples (blue font) were randomly selected, the selection being held constant, but presented in random order across queries. For details of descriptions and examples used, see Appendix I.

A3.4 Failed classifications

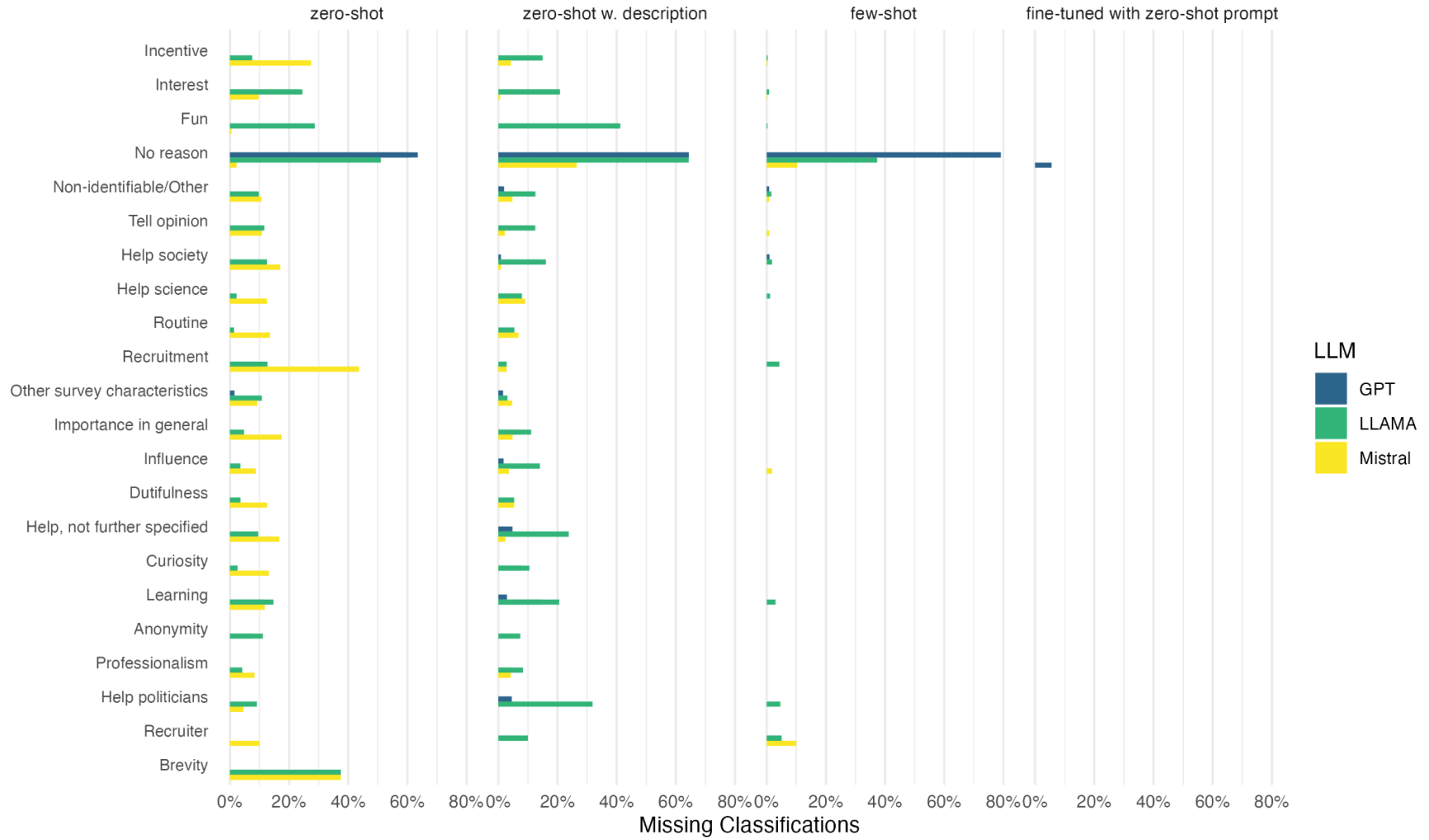


Figure A3.3: Proportion of missing classifications by category, LLM, and prompting approach.

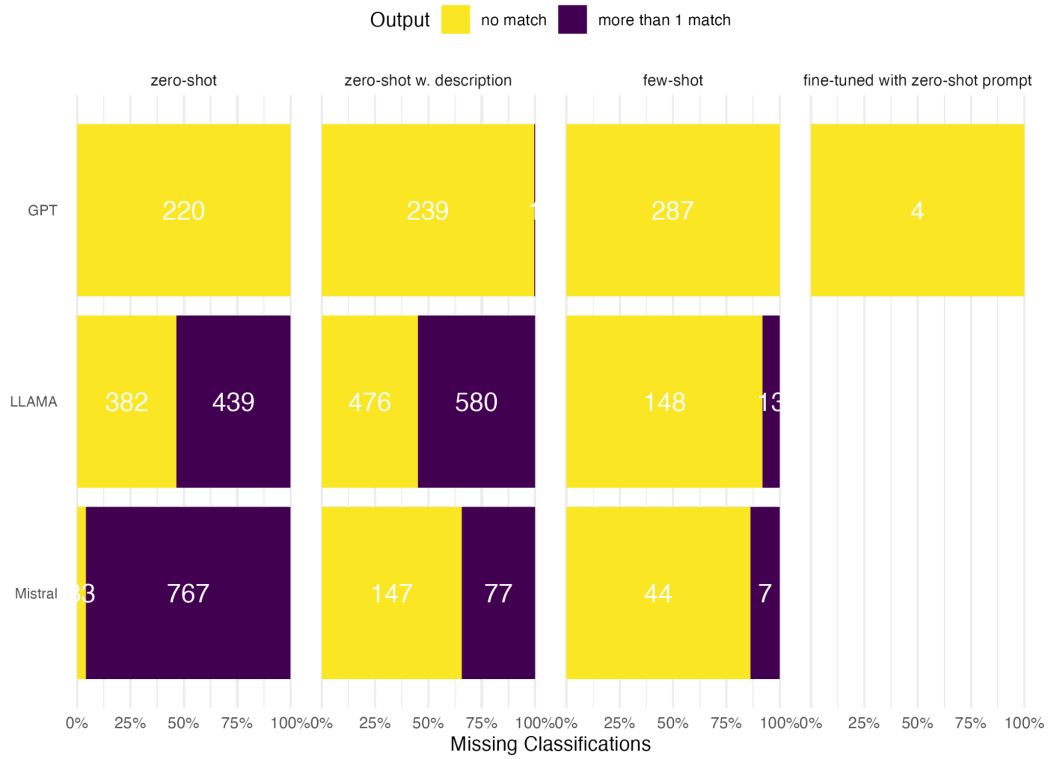


Figure A3.4: Distribution of outputs recorded as missing due to missing or ambiguous classifications.

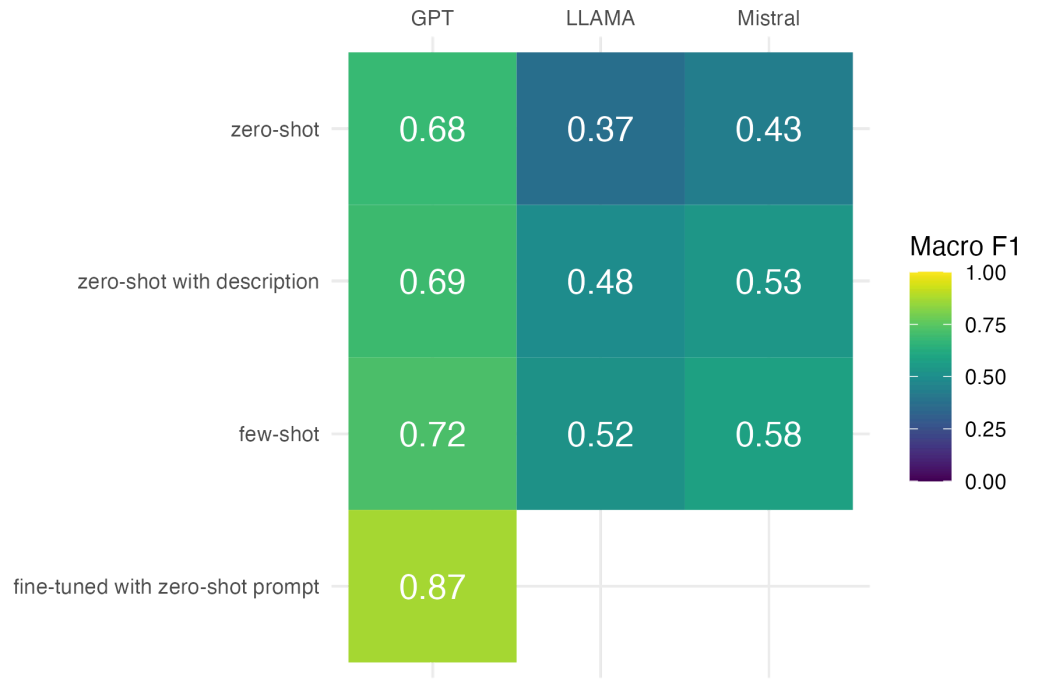


Figure A3.5: Macro F1 scores by LLM and prompting approach without missing values.

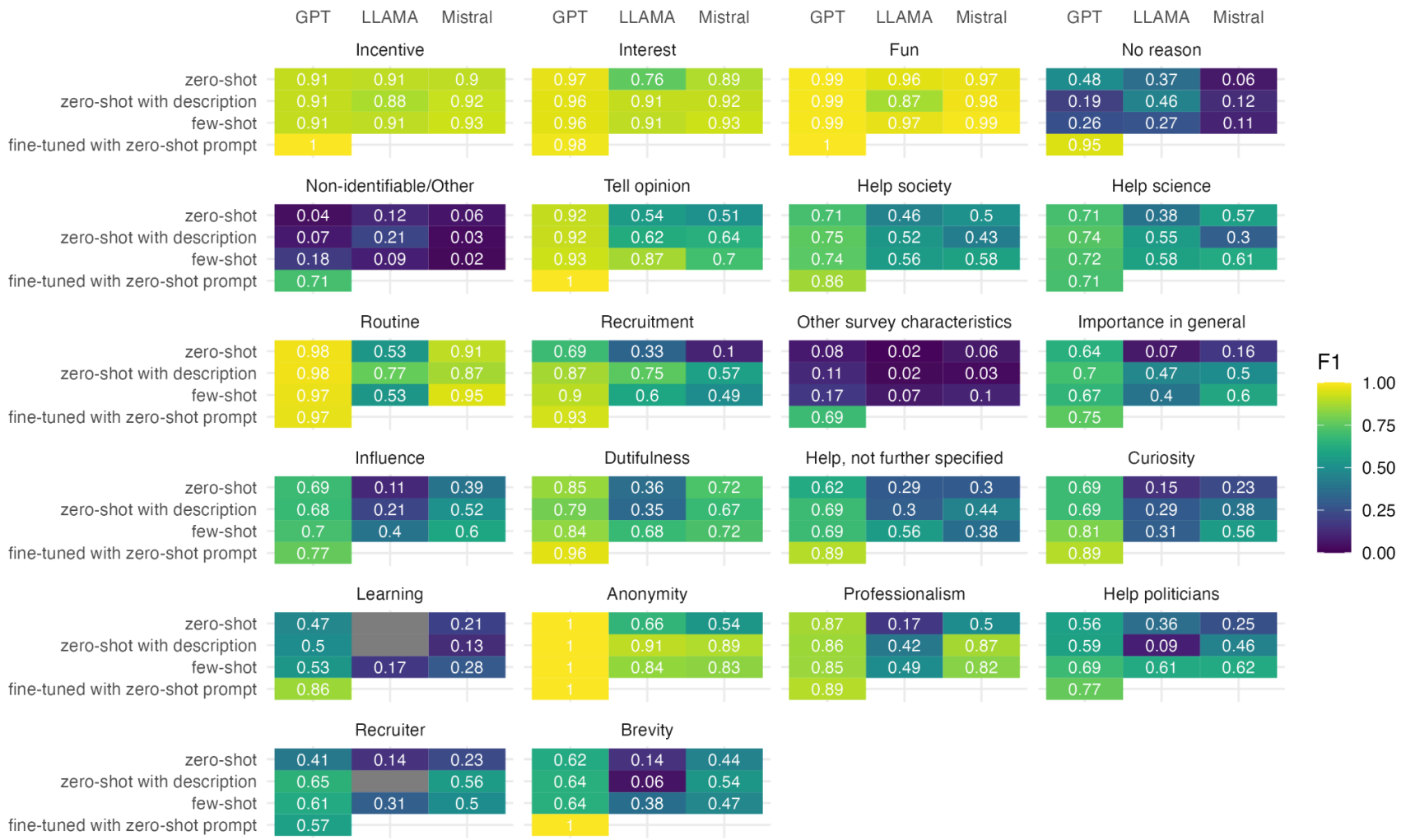


Figure A3.6: Per-category scores by LLM and prompting approach without missing values.

A3.5 Additional performance metrics

F1 Scores

F1 scores evaluate machine learning model accuracy by considering model precision and recall. Recall refers to the number of correct classifications for a given category (true positives) divided by the number of actual classifications for that category (true positives + false negatives), measuring the share of classifications for a category as coded by the human experts that were correctly classified by the LLM. Precision refers to the number of correct classifications for a given category (true positives) divided by the number of all classifications for that category (true + false positives), measuring the share of LLM classifications for a given category that were correct. The F1 score is defined as 2*precision*recall / (precision + recall), with a range of [0;1].

Macro F1 scores first calculate the F1 for each category separately and then average across these scores for a total model performance score, thereby taking into account unbalanced datasets, which may result in the LLM just assigning the modal category.

Weighted F1 scores adjust for class imbalance by taking a weighted average of per-category F1 scores based on the number of true instances in each category. They are calculated as the weighted harmonic mean of precision and recall. This helps balance the contributions of minority categories without disproportionately emphasizing them. They thus offer a compromise between macro F1 and accuracy (see below), balancing performance on both common and rare classes without over-penalizing a model that performs well on the modal category.

Accuracy

Accuracy describes the proportion of cases correctly classified out of all cases.

Intraclass Correlation Coefficient (ICC)

The ICC quantifies the agreement raters by evaluating the proportion of variance attributable to differences between raters, relative to the total variance. The type used here is ICC(2,1), which assumes a two-way random-effects model and measures agreement. Thus, it does not consider the human-generated classifications as a benchmark or ground truth, but simply as a different coder to compare the LLM to.

Cohen’s Kappa

Cohen's Kappa is a measure of agreement that takes into account the agreement occurring by chance, expressed as a proportion of the total possible improvement. It is particularly helpful for imbalanced datasets, as it quantifies model performance relative to a baseline of random chance, i.e., a naive classifier.

Metric	Approach	GPT-4o	Llama-3.2	Mistral NeMo
Weighted F1	*zero-shot*	0.72	0.40	0.46
	with descriptions	0.72	0.42	0.63
	few-shot	0.72	0.63	0.67
	fine-tuned	0.91		
Accuracy	*zero-shot*	0.81	0.55	0.60
	with descriptions	0.79	0.58	0.74
	few-shot	0.80	0.74	0.77
	fine-tuned	0.95		
ICC	*zero-shot*	0.83	0.37	0.61
	with descriptions	0.81	0.47	0.67
	few-shot	0.84	0.65	0.74
	fine-tuned	0.96		
Cohen's Kappa	*zero-shot*	0.76	0.47	0.53
	with descriptions	0.75	0.50	0.68
	few-shot	0.75	0.68	0.72
	fine-tuned	0.94		

Table A3.1: Classification performance metrics per LLM and prompting approach, compared to human labels.

A3.6 Confusion Matrices

GPT – zero-shot

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brevity	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
Curiosity	0	0	38	0	0	0	0	0	0	0	0	0	10	3	0	20	1	0	0	0	0	0
Dutifulness	0	0	0	48	0	0	0	1	1	0	0	0	0	0	0	6	0	0	0	0	1	0
Fun	0	0	0	0	376	0	0	0	0	0	0	0	1	1	0	2	5	0	0	0	0	0
Help politicians	0	0	0	0	0	22	1	1	1	0	0	7	2	1	0	21	0	0	0	0	0	0
Help science	0	0	0	0	0	0	79	2	11	12	0	3	1	0	0	26	0	1	0	0	0	0
Help society	0	0	0	0	0	0	0	96	7	5	0	9	1	1	0	40	0	0	0	0	0	0
Help, not further specified	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	2	0	0	0	1	0	0
Importance in general	0	0	0	0	0	0	1	10	0	42	0	0	0	0	0	12	4	0	0	0	0	0
Incentive	0	0	0	5	0	0	0	0	0	0	1523	0	0	0	0	3	0	0	1	0	0	0
Influence	0	0	0	0	0	0	0	0	0	1	0	38	1	0	0	12	0	0	0	0	0	1
Interest	0	0	0	0	0	0	0	1	0	1	0	0	1234	6	0	27	31	0	1	0	0	0
Learning	0	0	0	0	0	0	4	0	1	2	0	0	0	21	0	28	0	0	0	0	0	0
No reason	0	0	0	1	0	0	0	0	0	0	270	0	2	0	134	4	2	0	1	3	0	0
Non-identifiable/Other	0	0	0	1	0	0	0	0	0	0	3	0	0	0	6	5	1	0	1	0	0	0

Other survey characteristics	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	9	3	0	0	0	0	0
Professionalism	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	23	0	0	0	0
Recruiter	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	26	0	0
Recruitment	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4	41	1	0
Routine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	72	0
Tell opinion	0	0	0	0	0	0	2	1	0	0	0	0	0	1	0	36	1	0	0	0	0	231
NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	219	0	1	0	0	0	0	0

Table A3.2: Confusion matrix (actual vs. predicted categories) for GPT under zero-shot prompting.

GPT – with descriptions

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brevity	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0
Curiosity	0	0	38	0	0	0	1	0	0	1	0	0	12	1	0	18	1	0	0	0	0	0
Dutifulness	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Fun	0	0	0	0	376	0	0	0	0	0	0	0	0	1	0	2	4	0	0	0	0	0
Help politicians	0	0	0	0	0	17	1	1	0	0	0	2	0	0	0	15	1	0	0	0	0	0
Help science	0	0	0	0	0	0	72	0	5	7	0	2	0	0	0	20	0	0	0	0	0	0
Help society	0	0	0	0	0	1	0	104	7	4	0	8	1	1	0	39	0	0	0	0	0	0
Help, not further specified	0	0	0	0	0	0	0	0	24	0	0	3	0	0	0	3	0	0	0	0	0	0
Importance in general	0	0	0	0	0	0	2	3	0	47	0	1	0	0	0	16	2	0	0	0	0	0
Incentive	0	0	0	6	0	0	0	0	0	0	1522	0	0	0	0	1	0	0	1	0	0	0
Influence	0	0	0	0	0	3	0	2	0	0	0	40	1	0	0	12	0	0	0	0	0	4
Interest	0	0	0	0	0	0	4	1	0	1	0	0	1235	9	0	40	35	0	0	0	0	0
Learning	0	0	0	0	0	0	3	0	0	2	0	0	0	22	0	27	0	0	0	0	0	0
No reason	0	0	0	0	0	0	0	0	0	0	204	0	0	0	36	3	2	0	0	0	0	0

Non-identifiable/Other	0	0	0	1	0	0	0	0	0	0	71	0	2	0	91	16	2	0	0	0	0	0
Other survey characteristics	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	3	4	0	0	0	0	0
Professionalism	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	4	24	0	0	0	0
Recruiter	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	12	3	0	0
Recruitment	0	0	0	7	0	0	0	0	0	0	0	0	0	0	1	2	0	0	7	68	0	0
Routine	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	74	0
Tell opinion	0	0	0	0	0	0	2	0	1	1	0	0	0	0	0	33	0	0	0	0	0	228
NA	0	0	0	0	0	1	0	1	2	0	0	1	0	0	231	3	1	0	0	0	0	0

Table A3.3: Confusion matrix (actual vs. predicted categories) for GPT under zero-shot prompting with descriptions.

GPT – few-shot

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brevity	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0
Curiosity	0	0	38	0	0	0	0	0	0	0	0	0	5	2	2	9	0	0	0	0	0	0
Dutifulness	0	0	0	46	0	0	0	1	1	0	0	0	0	0	0	4	0	0	0	0	2	0
Fun	0	0	0	0	376	0	0	0	0	0	0	0	1	0	0	2	2	0	0	0	0	0
Help politicians	0	0	0	0	0	21	1	2	0	0	0	4	0	0	0	11	0	0	0	0	0	0
Help science	0	0	0	0	0	0	81	1	9	11	0	2	1	0	0	32	0	1	0	0	0	0
Help society	0	0	0	0	0	0	0	99	8	4	0	7	1	0	0	36	0	0	0	0	0	0
Help, not further specified	0	0	0	0	0	0	0	0	24	0	0	1	0	0	0	3	0	0	0	0	0	0
Importance in general	0	0	0	0	0	0	2	3	0	44	0	0	0	0	0	15	5	0	0	0	0	0
Incentive	0	0	0	5	0	0	0	0	0	0	1523	0	0	0	0	3	0	0	1	0	0	0
Influence	0	0	0	0	0	1	0	1	0	1	0	43	1	0	1	15	0	0	0	0	0	2
Interest	0	0	0	0	0	0	0	1	0	0	0	0	1193	3	0	21	27	0	0	0	0	0
Learning	0	0	0	0	0	0	3	2	0	3	0	0	1	28	0	33	0	0	0	0	0	1
No reason	0	0	0	0	0	0	0	0	0	0	250	0	2	0	50	3	4	0	0	0	0	0
Non-identifiable/Other	0	0	0	1	0	0	0	0	0	0	24	0	0	1	20	29	1	0	0	0	0	0

Other survey characteristics	0	0	0	0	0	0	0	0	0	0	0	0	46	0	0	7	11	0	0	0	0	0	0
Professionalism	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	6	23	0	0	0	0	0
Recruiter	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	11	3	0	0	0
Recruitment	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	68	1	0	0
Routine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	71	0	0
Tell opinion	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	29	0	0	0	0	0	229	0
NA	0	0	0	0	0	0	0	1	0	0	0	0	0	0	284	2	0	0	0	0	0	0	0

Table A3.4: Confusion matrix (actual vs. predicted categories) for GPT under few-shot prompting.

GPT – fine-tuned

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brevity	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Curiosity	0	0	8	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
Dutifulness	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fun	0	0	0	0	76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Help politicians	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
Help science	0	0	0	0	0	0	11	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
Help society	0	0	0	0	0	0	0	19	0	1	0	1	0	0	0	0	0	0	0	0	0	0
Help, not further specified	0	0	0	0	0	0	0	0	8	0	0	1	0	0	0	0	0	0	0	0	0	0
Importance in general	0	0	0	0	0	0	4	1	0	12	0	0	0	0	0	0	2	0	0	0	0	0
Incentive	0	0	0	1	0	0	0	0	0	0	360	0	0	0	0	0	0	0	0	0	0	0
Influence	0	0	0	0	0	0	0	1	0	0	0	10	0	0	0	3	0	0	0	0	0	0
Interest	0	0	0	0	0	0	1	1	0	0	0	0	248	0	1	1	0	0	1	0	0	0
Learning	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	1	0	0	0	0	0	0
No reason	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63	1	0	0	0	0	0	0
Non-identifiable/Other	0	0	0	0	0	0	2	1	1	0	0	0	1	1	4	34	1	0	0	0	0	0

Other survey characteristics	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	10	1	0	0	0	0
Professionalism	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
Recruiter	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0
Recruitment	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	0	0
Routine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	15	0
Tell opinion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47
NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0

Table A3.5: Confusion matrix (actual vs. predicted categories) for fine-tuned GPT with zero-shot prompt.

Llama – zero-shot

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	25	0	0	0	0	0	0	0	0	0	0	0	0	6	15	3	1	0	0	0	0	1
Brevity	0	3	0	1	0	0	0	0	0	0	15	0	0	0	7	2	3	0	0	4	2	0
Curiosity	0	0	34	0	0	2	2	1	0	2	26	2	287	8	23	18	6	0	1	0	0	7
Dutifulness	0	0	0	22	0	1	0	3	1	1	0	6	0	1	0	12	1	1	0	1	0	17
Fun	0	0	0	1	262	0	0	0	0	1	7	0	2	0	0	4	0	0	1	1	0	0
Help politicians	0	0	0	0	0	7	0	4	0	0	0	2	1	0	0	4	0	0	0	0	0	1
Help science	0	0	0	0	0	2	76	27	15	35	3	18	15	10	0	84	7	8	0	0	0	14
Help society	0	0	0	0	0	2	0	38	5	0	4	2	1	0	1	11	1	0	0	0	0	3
Help, not further specified	0	0	0	0	0	0	0	4	10	2	3	5	0	0	0	4	0	1	0	0	0	1
Importance in general	0	0	0	0	0	0	0	5	0	3	0	3	1	1	0	3	2	0	0	0	0	5
Incentive	0	0	0	2	0	0	0	0	0	0	1419	2	0	0	13	6	5	0	1	1	0	6
Influence	0	0	0	0	0	1	0	1	0	1	2	4	0	0	1	2	0	0	0	0	0	6
Interest	0	0	1	0	1	0	0	3	1	5	29	2	643	2	1	15	13	1	2	0	0	23
Learning	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	4	1	0

No reason	0	1	0	6	0	0	0	0	0	0	11	0	2	0	47	5	3	0	0	3	0	2
Non-identifiable/Other	0	0	0	1	1	0	0	0	1	1	13	0	3	1	32	18	2	0	2	0	0	3
Other survey characteristics	0	0	0	0	0	0	0	1	0	1	5	0	0	1	24	5	1	0	0	0	0	11
Professionalism	0	0	0	0	0	1	1	7	1	6	2	3	0	0	0	7	2	7	0	0	1	22
Recruiter	0	0	0	4	0	1	1	0	1	0	28	0	0	0	1	3	2	2	6	14	0	1
Recruitment	0	0	0	8	0	2	5	0	1	1	43	4	0	0	1	9	2	3	6	30	1	2
Routine	0	1	2	7	4	1	0	2	0	0	58	2	0	0	10	13	7	0	1	7	68	1
Tell opinion	0	0	0	0	0	0	1	1	2	1	0	0	0	0	0	3	0	0	0	0	0	80
NA	2	3	1	2	108	2	2	14	4	3	129	2	297	4	183	24	7	1	0	6	1	26

Table A3.6: Confusion matrix (actual vs. predicted categories) for Llama under zero-shot prompting.

Llama – with descriptions

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	27	0	0	0	0	0	1	0	0	0	2	0	0	0	0	3	1	0	0	0	0	0
Brevity	0	7	0	0	0	0	0	0	0	0	0	1	0	0	0	1	9	0	0	0	0	0
Curiosity	0	0	38	0	0	0	5	0	0	2	21	1	39	5	24	20	2	0	0	0	0	4
Dutifulness	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0
Fun	0	0	0	0	375	0	0	0	1	0	2	0	2	4	1	2	3	0	0	0	0	0
Help politicians	0	0	0	0	0	8	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	1
Help science	0	0	0	0	0	1	36	3	4	4	42	3	0	0	61	9	0	0	0	0	0	0
Help society	0	0	0	2	0	3	1	63	3	8	37	7	0	0	32	22	1	0	0	0	0	0
Help, not further specified	0	0	0	1	0	2	2	2	20	2	5	4	0	2	0	9	0	0	0	0	0	0
Importance in general	0	0	0	0	0	0	6	26	2	37	1	0	1	2	0	11	2	0	0	0	0	1
Incentive	0	0	0	6	0	1	0	0	0	0	1478	0	0	0	0	5	0	0	1	0	0	0
Influence	0	0	0	0	0	1	0	0	0	0	2	23	0	0	0	7	0	0	0	0	0	1
Interest	0	0	0	0	0	3	15	8	1	3	24	3	1201	13	10	62	28	0	1	0	0	6
Learning	0	0	0	0	0	0	8	1	0	1	3	1	0	4	3	7	0	0	0	0	0	1
No reason	0	1	0	0	0	0	0	2	1	0	6	2	0	0	19	6	1	0	0	0	0	4

Non-identifiable/Other	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	4	3	0	2	1	0	0
Other survey characteristics	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0
Professionalism	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	3	23	0	0	0	0
Recruiter	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	2	0	0	10	1	0	0
Recruitment	0	0	0	10	1	0	0	2	5	0	29	5	0	1	17	17	4	0	6	67	3	0
Routine	0	0	0	2	0	0	0	0	0	0	5	0	0	1	7	1	1	0	0	0	66	0
Tell opinion	0	0	0	0	0	3	4	3	2	3	63	4	0	2	88	45	3	0	0	0	0	209
NA	0	0	0	3	0	0	8	1	1	3	76	2	7	0	95	12	3	1	0	2	5	5

Table A3.7: Confusion matrix (actual vs. predicted categories) for Llama under zero-shot prompting with descriptions.

Llama – few-shot

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	27	0	0	0	0	0	0	0	0	0	2	0	0	0	7	0	0	0	1	0	0	0
Brevity	0	7	0	0	0	0	0	0	0	0	9	0	0	0	3	0	8	0	0	2	0	0
Curiosity	0	0	38	0	1	0	0	0	0	0	58	2	16	15	59	11	3	0	0	0	0	2
Dutifulness	0	0	0	35	0	0	0	1	1	0	1	2	0	0	3	4	0	0	0	0	0	0
Fun	0	0	0	0	374	0	0	0	0	0	14	0	5	0	3	2	0	0	0	0	0	1
Help politicians	0	0	0	0	0	14	0	0	0	0	0	3	2	0	1	4	0	0	0	0	0	1
Help science	0	0	0	0	0	2	80	4	8	15	0	4	2	3	6	57	3	2	0	0	0	3
Help society	0	0	0	0	0	3	2	94	6	19	1	13	6	2	8	65	3	0	0	0	0	3
Help, not further specified	0	0	0	0	0	0	0	0	21	1	3	2	0	0	2	4	0	0	0	0	0	0
Importance in general	0	0	0	1	0	0	0	7	1	22	0	3	0	0	6	6	0	0	0	0	0	2
Incentive	0	0	0	4	0	0	1	0	0	0	1505	0	0	0	2	2	0	0	1	0	0	0
Influence	0	0	0	0	0	2	1	0	0	0	1	21	2	1	5	9	3	0	0	0	0	4
Interest	0	0	0	0	0	0	0	1	1	1	36	0	1107	1	3	10	19	0	0	1	0	10
Learning	0	0	0	0	0	0	2	0	0	0	0	0	0	4	3	4	0	1	0	0	0	1
No reason	0	1	0	2	0	0	0	0	0	1	36	0	0	0	43	6	3	0	0	0	0	0
Non-identifiable/Other	0	0	0	1	0	0	0	0	0	0	25	1	3	4	20	15	4	0	1	0	0	1
Other survey characteristics	0	0	0	0	0	0	0	0	1	0	3	1	100	1	17	15	7	0	0	0	0	0

Professionalism	0	0	0	1	0	0	0	1	1	4	2	1	0	2	3	11	5	19	1	0	0	2
Recruiter	0	0	0	3	0	0	1	0	0	0	2	0	0	0	5	3	0	2	12	30	0	0
Recruitment	0	0	0	3	0	0	0	0	0	0	1	0	0	0	4	1	0	0	2	34	0	1
Routine	0	0	0	6	0	0	0	2	0	0	93	0	0	0	17	7	6	0	1	1	74	0
Tell opinion	0	0	0	0	0	0	0	1	2	0	0	4	0	0	5	15	1	0	0	0	0	201
NA	0	0	0	0	1	1	1	1	0	0	5	0	9	1	134	4	0	0	1	3	0	0

Table A3.8: Confusion matrix (actual vs. predicted categories) for Llama under few-shot prompting.

Mistral – zero-shot

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	27	0	0	2	0	0	0	4	1	2	21	1	1	1	1	8	3	0	1	0	0	0
Brevity	0	4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	7	0	0	1	0	0
Curiosity	0	0	31	1	1	3	4	9	2	7	40	3	68	6	12	35	10	0	0	1	0	1
Dutifulness	0	0	0	31	0	0	0	0	0	0	0	1	0	0	1	2	1	1	0	0	0	0
Fun	0	0	0	0	372	0	0	0	0	0	7	0	3	1	0	4	5	0	1	0	0	0
Help politicians	0	0	0	0	0	5	0	0	0	0	0	0	1	0	10	3	0	0	0	0	0	0
Help science	0	0	0	0	0	3	52	4	5	6	2	0	3	1	1	26	0	2	0	0	0	1
Help society	0	0	0	2	0	1	0	46	4	6	3	4	0	0	1	23	0	0	0	0	0	0
Help, not further specified	0	0	0	1	0	2	0	1	9	0	1	6	0	1	0	2	1	0	1	1	0	0
Importance in general	0	0	0	0	0	1	1	9	3	8	5	4	0	4	0	6	3	0	0	0	0	1
Incentive	0	0	0	6	0	0	2	0	0	0	1087	1	0	0	0	5	0	0	1	0	0	0
Influence	0	0	0	1	0	2	0	1	0	1	5	19	0	0	3	11	0	0	0	0	0	2
Interest	0	0	2	1	1	2	7	5	3	11	74	1	1049	7	9	41	18	0	3	0	0	2
Learning	0	0	0	1	0	0	7	4	1	0	11	1	2	8	0	10	1	0	0	1	0	1
No reason	0	1	0	0	0	0	0	1	0	0	2	0	0	0	12	3	1	0	1	2	0	0
Non-identifiable/Other	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	7	0	0	1	0	0	0
Other survey	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0

characteristics																						
Professionalism	0	0	0	0	0	0	2	3	0	3	8	4	1	0	0	6	1	17	1	0	0	0
Recruiter	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	6	27	0	0
Recruitment	0	0	0	1	0	1	0	0	1	0	14	1	0	0	0	7	3	0	2	4	3	0
Routine	0	0	0	1	0	0	0	0	0	0	4	0	0	0	0	1	1	0	0	2	61	0
Tell opinion	0	0	0	0	0	1	2	6	6	8	20	5	2	1	298	28	2	2	0	0	0	199
NA	0	3	5	7	2	1	11	19	7	11	493	5	121	4	8	27	6	2	2	31	10	25

Table A3.9: Confusion matrix (actual vs. predicted categories) for Mistral under zero-shot prompting.

Mistral – with descriptions

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	27	0	0	0	0	0	1	0	0	0	2	0	0	0	0	3	1	0	0	0	0	0
Brevity	0	7	0	0	0	0	0	0	0	0	0	1	0	0	0	1	9	0	0	0	0	0
Curiosity	0	0	38	0	0	0	5	0	0	2	21	1	39	5	24	20	2	0	0	0	0	4
Dutifulness	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0
Fun	0	0	0	0	375	0	0	0	1	0	2	0	2	4	1	2	3	0	0	0	0	0
Help politicians	0	0	0	0	0	8	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	1
Help science	0	0	0	0	0	1	36	3	4	4	42	3	0	0	61	9	0	0	0	0	0	0
Help society	0	0	0	2	0	3	1	63	3	8	37	7	0	0	32	22	1	0	0	0	0	0
Help, not further specified	0	0	0	1	0	2	2	2	20	2	5	4	0	2	0	9	0	0	0	0	0	0
Importance in general	0	0	0	0	0	0	6	26	2	37	1	0	1	2	0	11	2	0	0	0	0	1
Incentive	0	0	0	6	0	1	0	0	0	0	1478	0	0	0	0	5	0	0	1	0	0	0
Influence	0	0	0	0	0	1	0	0	0	0	2	23	0	0	0	7	0	0	0	0	0	1
Interest	0	0	0	0	0	3	15	8	1	3	24	3	1201	13	10	62	28	0	1	0	0	6
Learning	0	0	0	0	0	0	8	1	0	1	3	1	0	4	3	7	0	0	0	0	0	1
No reason	0	1	0	0	0	0	0	2	1	0	6	2	0	0	19	6	1	0	0	0	0	4
Non-identifiable/Other	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	4	3	0	2	1	0	0
Other survey	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0

characteristics																						
Professionalism	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	3	23	0	0	0	0
Recruiter	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	2	0	0	10	1	0	0
Recruitment	0	0	0	10	1	0	0	2	5	0	29	5	0	1	17	17	4	0	6	67	3	0
Routine	0	0	0	2	0	0	0	0	0	0	5	0	0	1	7	1	1	0	0	0	66	0
Tell opinion	0	0	0	0	0	3	4	3	2	3	63	4	0	2	88	45	3	0	0	0	0	209
NA	0	0	0	3	0	0	8	1	1	3	76	2	7	0	95	12	3	1	0	2	5	5

Table A3.10: Confusion matrix (actual vs. predicted categories) for Mistral under zero-shot prompting with descriptions.

Mistral – few-shot

R: Prediction / C: Reference	Anonymity	Brevity	Curiosity	Dutifulness	Fun	Help politicians	Help science	Help society	Help, not further specified	Importance in general	Incentive	Influence	Interest	Learning	No reason	Non-identifiable/ Other	Other survey characteristics	Professionalism	Recruiter	Recruitment	Routine	Tell opinion
Anonymity	27	0	0	0	0	0	0	0	0	0	7	0	0	0	1	3	0	0	0	0	0	0
Brevity	0	8	0	0	0	0	0	0	0	0	0	0	0	0	6	1	11	0	0	0	0	0
Curiosity	0	0	38	0	1	0	3	0	0	1	17	0	15	10	4	9	0	0	0	0	0	0
Dutifulness	0	0	0	33	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Fun	0	0	0	0	375	0	0	0	0	0	0	0	2	0	0	0	5	0	0	0	0	0
Help politicians	0	0	0	0	0	12	0	0	0	0	1	1	1	0	0	2	0	0	0	0	0	0
Help science	0	0	0	0	0	3	74	3	7	12	2	1	2	0	1	49	1	1	0	0	0	0
Help society	0	0	0	0	0	4	0	94	10	6	2	12	2	1	35	44	1	0	0	0	0	0
Help, not further specified	0	0	0	2	0	0	4	6	21	0	18	1	0	0	12	5	0	0	0	0	0	0
Importance in general	0	0	0	0	0	0	1	6	0	42	9	2	0	1	0	12	3	0	0	0	0	0
Incentive	0	0	0	7	0	0	0	0	0	0	1572	0	0	0	18	7	0	0	1	0	0	0
Influence	0	0	0	1	0	3	0	0	0	0	1	31	0	0	0	10	0	0	0	0	0	2
Interest	0	0	0	0	0	0	0	1	0	0	28	0	1162	5	5	28	25	0	0	0	0	0
Learning	0	0	0	0	0	0	2	1	0	0	0	0	0	8	4	8	0	0	0	0	0	0
No reason	0	0	0	0	0	0	0	0	0	0	2	1	0	0	19	9	3	0	0	0	0	0
Non-identifiable/Other	0	0	0	0	0	0	0	0	0	0	28	0	2	3	50	4	1	0	0	0	0	0
Other survey	0	0	0	0	0	0	0	0	0	0	2	0	62	2	20	13	9	0	0	0	0	0

characteristics																						
Professionalism	0	0	0	0	0	0	2	0	0	0	2	0	0	0	0	1	4	23	0	0	0	0
Recruiter	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	7	2	0	0
Recruitment	0	0	0	13	0	0	0	0	3	0	95	4	1	0	0	11	2	0	10	69	2	1
Routine	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	3	0	0	0	0	72	0
Tell opinion	0	0	0	0	0	0	2	0	1	2	3	3	1	4	147	32	0	0	0	0	0	227
NA	0	0	0	0	0	0	0	0	0	0	5	1	2	0	37	2	0	0	2	0	0	2

Table A3.11: Confusion matrix (actual vs. predicted categories) for Mistral under few-shot prompting.

A3.7 Fine-Tuning Performance

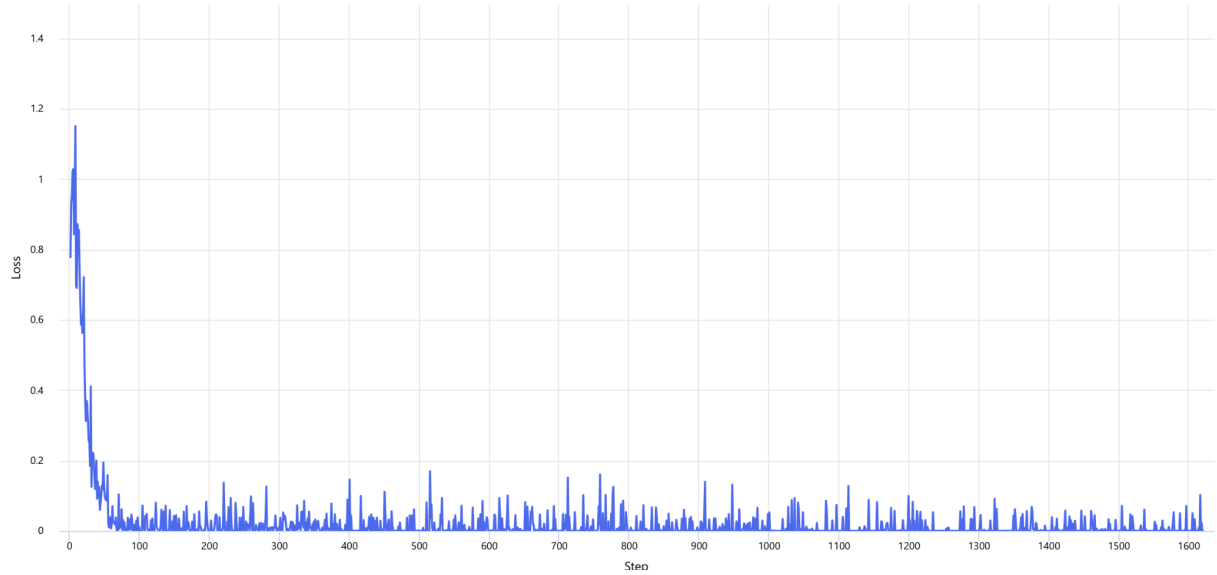


Figure A3.7: Loss curve during fine-tuning.
*Note: steps = n_examples/batch_size*n_epoch. Epochs describe the number of iterations through the data, and batch size the number of examples used in a single training pass.*

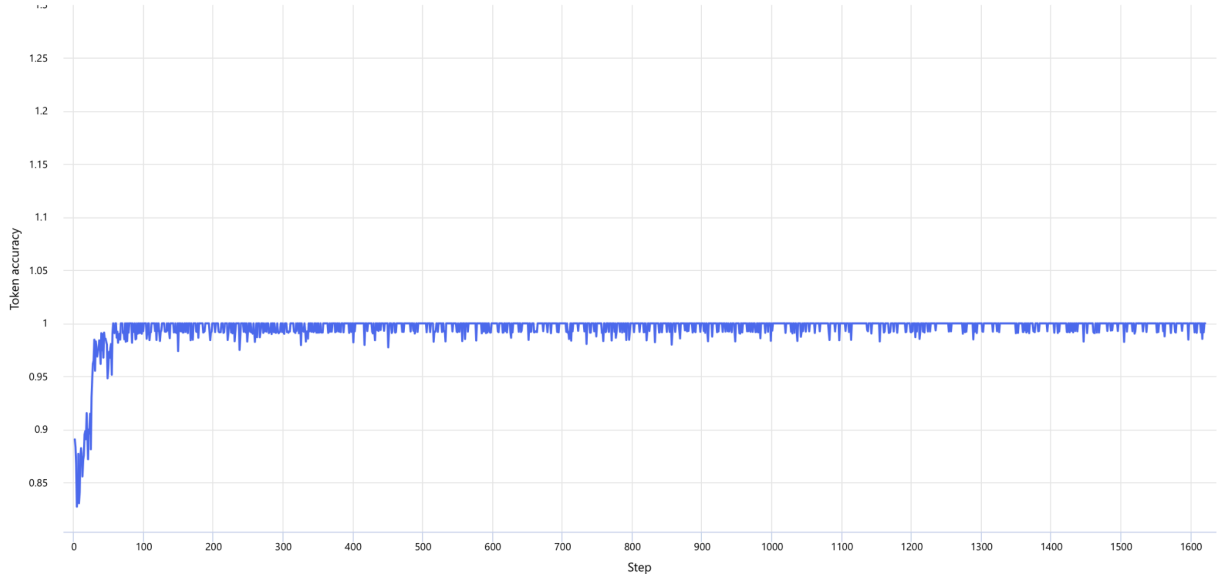


Figure A3.8: Mean token accuracy achieved during fine-tuning.