Who Counts? The Potentials and Pitfalls of Using LLMs in Survey Research

Leah von der Heyde
Social Data Science & AI Lab, LMU Munich & Munich Center for Machine Learning
l.heyde@lmu.de

Abstract

The integration of large language models (LLMs) into surveys presents opportunities for mitigating ongoing challenges regarding coverage, sampling, measurement, and nonresponse, all the while making survey research more efficient. However, LLMs can also introduce new challenges. As LLMs have only emerged rather recently as a potential tool in the survey methodologists' toolbox, how their use can improve versus worsen survey data quality has not been systematically investigated. In this paper, I present an overview of the potential applications of LLMs in survey research, and highlight their possible pitfalls. I identify three main roles LLMs can play in the survey research process: they can act as research assistants, interviewers, and respondents, with potential applications in all stages of the survey research process. I also discuss how LLM training, alignment, and model architectures, as well as research design choices can inhibit survey data quality, concluding that LLM-induced errors need to be investigated both methodologically and empirically, and that, short of mitigating ensuing biases, humans need to remain in the loop for making all humans count.

1 Introduction

Obtaining high-quality data is crucial for making valid inferences about people's attitudes and behaviors, be it in the context of research or policymaking. Therefore, safeguarding data quality lies at the heart of survey methodology. While survey researchers continue to face challenges in coverage, sampling, measurement, and nonresponse, the digitalization of society has allowed them to explore a variety of new options for collecting or supplementing survey data. Most recently, large language models (LLMs) have been the target of large hopes for alleviating existing challenges in survey research. However, LLMs are based on Internet data, which is not generated with the primary goal of producing population-level estimates and correlates (Groves & Lyberg, 2010; Salganik, 2019). As a result, LLMs may come with similar pitfalls as other digital data sources regarding inferences about human attitudes and behavior, which can jeopardize data quality. For example, coverage error might arise due to the digital divide (Lutz, 2019) and the selective corpora used for building LLMs. Thus, *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, LLMs' idiosyncratic data-generating processes put into question *who does the counting* – researchers, data curators, data annotators, algorithms? – and *what is being counted and how it is being counted*: Measurement error might arise, for example, because online behaviors are not always valid indicators of attitudes or behaviors (e.g., Jungherr et al., 2017), and LLMs' output of most likely next words in a sentence might not be valid indicator of social science concepts. Thus, LLMs not only have the potential to mitigate, but also to amplify existing biases regarding the understanding of different populations and constructs of interest. *Who counts* in LLM-augmented survey research continues to hinge on fundamental questions of representation and measurement.

With this paper, I present an overview of potential applications of LLMs in survey research based on a narrative review of existing literature, and discuss possible pitfalls in relation to

data quality. Providing a resource for survey methodologists as well as practitioners, this work contributes to the development of LLM-augmented survey research.

2 Potential Applications of LLMs in the Survey Research Process

In general, three main potential application areas for LLMs within the survey research process can be identified (Bail, 2024; Kreuter, 2025) – LLMs acting as research assistants, LLMs acting as interviewers, and LLMs acting as respondents. As such, use cases for LLMs are conceivable in all stages of the survey research process, where they could possibly help address errors impacting data quality while being more efficient than humans. This section provides an overview of the potential applications of LLMs in the survey research process before, during, and after data collection. For a graphic overview, see Figure A1 in the Appendix.

2.1 Pre-data collection

In the **questionnaire design** phase, LLMs have some potential to reduce human-induced validity issues. Built for creative text generation, possible applications of LLMs include 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, experimental vignettes or images (Bail, 2024; Demszky et al., 2023; Sarstedt et al., 2024), or entire questionnaires. Here, they likely need more instruction compared to a seasoned survey methodologist. LLMs might be more helpful when evaluating existing questions (Hommel, 2023; Olivos & Liu, 2024) and suggesting improvements (Jacobsen et al., 2025; Thirunavukarasu & O'Logbon, 2024) regarding readability, social desirability, leading or biased wording, or different literacy levels and cultural contexts. Another major 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; Metheney & Yehle, 2024), which usually features two independent human translations and a human adjudicator in case of disagreements.

During **pre-testing**, LLMs have the potential to mitigate measurement error by identifying patterns in pilot surveys (Kreuter, 2025). With even deeper integration, LLMs could act as virtual or simulated respondents (see subsection 2.2). 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 somewhat aid in defining the target population and suggest appropriate sampling frames based on the research questions and analyses of previous surveys and research papers. They could both review existing sampling plans, highlighting potential biases or limitations and making suggestions for improvement, and summarize best practices and recommend new sampling designs. 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 media formats, 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.

2.2 Data Collection

LLMs have major potential in the data collection phase of survey research. Here, they could **augment interviewing** by dynamically adapting the questionnaire based on previously given responses, for example through probing questions (Barari et al., 2024; Geisen, 2024) or by detecting acute respondent burden, inattentiveness, item nonresponse, or imminent

breakoff and deploying real-time mitigation strategies. 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). The automatic creation and real-time fielding of new survey items from open-ended responses would allow for the standardized measurement of relevant emerging topics (Velez, 2025). The increased relevance and responsiveness of such dynamic surveys could help improve survey engagement, 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 (Grossmann et al., 2023; Barari et al., 2024; Lerner, 2024). 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 in web-based surveys. Telephone surveys with LLM-based interviewers (e.g., Lang & Eskenazi, 2025; Leybzon et al., 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 (Jansen et al., 2023), and (in the audio-based formats) ease participation by visually or reading-impaired persons.
However, 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 highly prominent application of LLMs is that of simulating respondents through LLM-based **synthetic samples**, which can be relevant for several stages of the survey research process.
To create synthetic samples, an LLM is prompted to generate an artificial dataset of survey responses to the question(s) of interest. LLMs can be provided with individual-level information about humans, for example socio-demographic and attitudinal information collected in surveys, and are then asked to respond to survey questions from the respective person's perspective (e.g., Argyle et al., 2023; Bisbee et al., 2024; Simmons & Hare, 2023). This persona-based approach allows researchers to approximate specific target populations, simulating a vast array of individual positionalities and response distributions. This has been argued to address generalizability concerns (Grossmann et al., 2023) and help address coverage and sampling errors, potentially making LLM-based synthetic samples better-suited for social science research than the convenience samples used in many studies (Bail, 2024).
Synthetic samples could be used to supplement existing survey data during instrument development, data collection, and processing, and could potentially aid in reducing four major components of total survey error – coverage, sampling, measurement, and nonresponse. During pre-testing (e.g., Webb, 2024), they can help save 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). Furthermore, (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. 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; Jansen et al., 2023, but see subsection 3.3). Due to such considerations, 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). While the outright replaceability of human respondents during data collection is contested (see Agnew et al., 2024, for a review of positions), synthetic samples could be employed for imputing missing data due to unit- or

item-nonresponse after data collection – for example on sensitive topics or with hard-to-reach populations (Grossmann et al., 2023; Jansen et al., 2023; Kalinin, 2023; Holtdirk et al., 2024) – or by generating data for single items previously unasked (Kim & Lee, 2023).

2.3 Post-data collection

Upon the completion of data collection, multimodal LLMs (LLMs able process and/or generate image, audio, or video data) have major potential in helping with **data processing**. This includes digitizing survey data, for example by transcribing audio data from in-person, phone, or web-based interviews (Revilla et al., 2025; Tewari & Hosein, 2024) or transforming scans of paper-based (mail) questionnaires into tabular data with optical character recognition. While this is yet to be explored, it could 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 donations of individual-level digital behavioral data. 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, for example based on 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 make the coding text, image, or audio data much more efficient (see Ziems et al., 2024, for a systematic review). In the survey context, such data can, for example, come from open-ended responses (e.g., von der Heyde et al., 2025), social media, voice-based responses, or surveys asking respondents to upload pictures of their surroundings (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 (Bail, 2024; Demszky et al., 2023; Gilardi et al., 2023; Ahnert et al., 2025; Cerina & Duch, 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.

Smaller gains in reducing time and processing error could be achieved by having LLMs generate standardized and easy-to-use variable labels for datasets or entire codebooks. Given information about the data structure, they can also assist in writing code for data processing and analysis in a variety of programming languages, such as R or Python. 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 (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; 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. 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 Data Quality Challenges in LLM-Augmented Survey Research

Not long after their wider release, LLMs' output was found to be biased – also in relation to aspects of central importance for researching human attitudes and behaviors: LLMs exhibit cultural and psychological biases, including a tendency towards reflecting or assuming WEIRD (Western, Educated, Industrialized, Rich, and Democratic) norms and traits (e.g., Bianchi et al., 2023; Havaldar et al., 2023; Johnson et al., 2022; Masoud et al., 2025; Palacios Barea et al., 2023; Ramezani & Xu, 2023; 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). 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; Wang et al., 2024a). 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. This section discusses these potential error sources and their consequences in more depth. For a tabular overview of potentials and pitfalls, see Table A1 in the Appendix.

3.1 Training and alignment

While LLM training data corpora contain a large selection of human-generated text, this selection is not balanced. The corpora likely 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". Such biases can be especially detrimental when generating or processing survey data, but also when it comes to adapting or administering questions with LLMs. Although models based on biased data *can* sometimes make correct predictions, those predictions are neither reliable nor valid, as the Google Flu Trends example (Lazer et al., 2014) shows. With human attitudes and behaviors fluctuating, knowledge of what data was used to arrive at the predictions, and how those sources might impact the outcome, is crucial for social science inference.

Regarding the **digital divide**, bias is potentially introduced at several levels: First, there are cross-national differences in Internet access and behavior (Union, 2022; Schumacher & Kent, 2023). Although global Internet penetration rates are by now high, people without Internet access almost exclusively live in non-WEIRD countries (Crockett & Messeri, 2023; 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; Hoffmann et al., 2015; Shaw & Hargittai, 2018; Tucker et al., 2018; Hargittai, 2020; Kim et al., 2021) and interact with the cross-national differences (Schumacher & Kent, 2023). Because

of these disparities, even if Internet text was randomly selected for LLM training, certain populations and subgroups would be systematically under- or overrepresented (see also "Big Data Error": Amaya et al., 2020). 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; von der Heyde et al., 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; Roberts, 2022; Kuntz & Silva, 2023; 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 (Hargittai, 2015; Bradley et al., 2021). 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; Kuntz & Silva, 2023). The explicit and implicit attitudinal and behavioral biases expressed by the authors of the texts in the selected datasets not only risk getting encoded, but being disproportionately amplified in LLMs (Bender et al., 2021). For example, Heseltine & Clemm von Hohenberg (2024) found that LLMs performed worse when labeling non-English political texts (see also von der Heyde et al., 2025, for a survey research application). 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). In part, such issues might be tackled by fine-tuning [1] LLMs with relevant social media or survey data (e.g., Ahnert et al., 2025; Holtdirk et al., 2024). Nevertheless, 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 its affordances and algorithms. Individuals might use certain expressions to make their content more engaging (Buschek & Thorp), 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

[1] Further training (optimizing) of LLMs with a dataset of input-output pairs for the specific use case.

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 (e.g., von der Heyde et al., 2024). 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, resulting in faulty measurement. Representational issues could arise if, e.g., training data cutoffs preclude the realignment of political ideology and attitudes. For example, 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).

3.2 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 [2] 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 (Jansen et al., 2023). It is likely that LLMs are better-positioned for tasks closer to their original purpose of text generation and processing – that is, as a research assistant before (instrument development) and after (data labeling) data collection. 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 current 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; 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 3.1 might inhibit the generalizability of these studies' findings, and survey methodologists need to investigate whether LLMs, or a specific LLM (see subsection 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 (e.g., Ball et al., 2025), possibly as a result of explicit or implicit biases in the training data as well as the tokenizers. 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,

[2] Tasks that are easy for an LLM to complete because it has been provided with more training data enabling it to fulfill the task, like common logical problems or English-language text.

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., opinion prediction or 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. 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 (Wang et al., 2024b). 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 features**, especially when using them for data generation. 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 not being captured by LLM output (Grossmann et al., 2023; Demszky et al., 2023), although they do exist among humans. Similar to social desirability in surveys, this "machine desirability" can negatively impact measurement. Relatedly, due to their preprogrammed goal of agreeableness, LLMs might have a tendency for acquiescence bias (Bail, 2024; Dentella et al., 2023).

3.3 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[3] tend to reflect more lower-income, conservative views, whereas instruction-tuned GPT models have a liberal elite bias (Dillion et al., 2023). 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 (proprietary) LLMs, which may not always carry performance improvements for the specific task at hand, impacting reliability.

Further, the variability of model **hyperparameters** potentially inhibits data quality of LLM-augmented survey research. For example, while the amount of randomness in LLM outputs can be reduced by lowering the temperature, thereby increasing reliability by forcing the

[3]LLMs that have not undergone alignment based on human feedback.

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 percent 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.

Finally, LLMs' sensitivity to **prompt wording** (e.g., Bisbee et al., 2024; Gui & Toubia, 2023; Pezeshkpour & Hruschka, 2024) poses another challenge for data quality in LLM-augmented survey research. Both 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 & 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).

4 Conclusion

As this review has demonstrated, 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. It is not clear how exactly their errors play out in applied research. Traditionally, error frameworks have been developed and used for identifying and quantifying the errors that can arise at different steps of the survey research process. As has been the case for other new data sources (see Daikeler et al., 2024, for a review), such frameworks need to be adapted to integrate LLMs (see section 5).

At the very least, the potential existence of such errors put into question the generalizability of singular studies showcasing the successful application of LLMs to different survey research tasks. Many of these studies have been conducted in high-resource contexts, and LLMs were not a priori designed for survey research. Additionally, the relevance of LLM strengths and weaknesses varies across tasks in the survey research process. Thus, any survey-related application of LLMs needs to be evaluated for the specific task and context it is to be employed in.

As has become evident through this paper, LLMs have the potential to revolutionize survey research – for better or for worse. 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. With the necessary knowledge about their potentials and limitations, as well as human supervision and validation, it is possible that LLMs can be integrated with more tried and tested tools to provide a better picture of how societies think and act.

5 Limitations

While this overview has outlined the potentials and pitfalls of using LLMs in survey research, several limitations call for future research:

Systematic review of empirical findings. As a narrative review, this paper is mainly conceptual and only presents empirical findings as representative examples of potential

applications. As such, it aims to encourage methodologists and practitioners to engage more deeply with the specific aspects discussed. Although this is a comprehensive overview, the studies featured in it were not selected systematically. A more systematic literature review could quantify and synthesize the state of empirical implementation of the different potential applications and their respective findings, giving detailed insight into the conditions of success and failure. Such a review could also highlight gaps in methodological and empirical research, inviting further examination.

Concrete guidelines. Relatedly, developing an overview of approaches and best practices and error mitigation strategies for ensuring data quality for different kinds of survey-related LLM applications is a task for future work. For these efforts, it would be valuable if survey researchers engaged with the natural language processing (NLP) community, which has been working on understanding and improving LLMs from a technical point of view. Developing a shared vocabulary (see, e.g., Simmons & Hare, 2023) can help create synergies for improving data quality for and of LLMs. Beyond this, including a broader range of stakeholders, for example, pollsters, interviewers, respondents, and annotators, can highlight practical needs and ethical concerns for LLM integration. The considerations discussed in this paper can offer a starting point for such a dialog.

Adaptation of TSE framework. While the present paper separately discusses how LLMs can improve or worsen data quality in the survey research process, future research should develop a standardized and unified error framework that allows researchers to quantify biases, consider tradeoffs, and have specific standards for acceptable performance. Several of the error sources previously identified for multinational surveys (Pennell et al., 2017), Big Data (Amaya et al., 2020), and digital trace data (Sen et al., 2021) can likely be transferred to LLM-assisted survey research. In addition, as this review has highlighted, LLMs' idiosyncratic features and research designs can also introduce new error sources, calling for yet another adaptation of the Total Survey Error (TSE) framework (Groves et al., 2009; Groves & Lyberg, 2010) to the LLM-augmented reality of survey research, organized either along applications or along error sources. Integrating traditional, previously identified, and LLM-specific errors into a unified framework would be a helpful contribution to both the survey research community and the NLP community that is developing LLMs, which would be provided with guidance for identifying biases, contributing to efforts to mitigate them.

References

Divya Mani Adhikari, Vikram Kamath Cannanure, Alexander Hartland, and Ingmar Weber. Exploring LLMs for Automated Pre-Testing of Cross-Cultural Surveys, 2025. URL http://arxiv.org/abs/2501.05985v1.

William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. The Illusion of Artificial Inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi: 10.1145/3613904.3642703. URL https://doi.org/10.1145/3613904.3642703. event-place: Honolulu, HI, USA.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. Place: Honolulu, Hawaii, USA.

Georg Ahnert, Max Pellert, David Garcia, and Markus Strohmaier. 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, 2025. doi: 10.1609/icwsm.v19i1.35801. URL https://doi.org/10.1609/icwsm.v19i1.35801.

Ashley Amaya, Paul P Biemer, and David Kinyon. Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, 8(1):89–

119, February 2020. ISSN 2325-0984, 2325-0992. doi: 10.1093/jssam/smz056. URL https://academic.oup.com/jssam/article/8/1/89/5728725.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, pp. 1–15, February 2023. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2023.2. URL https://www.cambridge.org/core/product/identifier/S1047198723000025/type/journal_article.

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. Which Humans?, September 2023. URL https://osf.io/5b26t.

Christopher A. Bail. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, May 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2314021121. URL https://pnas.org/doi/10.1073/pnas.2314021121.

Sarah Ball, Simeon Allmendinger, Frauke Kreuter, and Niklas Kühl. Human Preferences in Large Language Model Latent Space: A Technical Analysis on the Reliability of Synthetic Data in Voting Outcome Prediction, February 2025. URL http://arxiv.org/abs/2502.16280. arXiv:2502.16280 [cs].

Soubhik Barari, Zoe Slowinski, Natalie Wang, Jarret Angbazo, Brandon Sepulvado, Leah Christian, and Elizabeth Dean. Generative AI Can Enhance Survey Interviews. Technical report, NORC at the University of Chicago, November 2024. URL https://www.norc.org/research/library/generative-ai-can-enhance-survey-interviews.html.

Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. GermanPartiesQA: Benchmarking Commercial Large Language Models for Political Bias and Sycophancy, July 2024. URL http://arxiv.org/abs/2407.18008. arXiv:2407.18008 [cs].

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, Chicago IL USA, June 2023. ACM. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3594095. URL https://dl.acm.org/doi/10.1145/3593013.3594095.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, pp. 1–16, May 2024. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2024.5. URL https://www.cambridge.org/core/journals/political-analysis/article/synthetic-replacements-for-human-survey-data-the-perils-of-large-language-models/B92267DC26195C7F36E63EA04A47D2FE.

Grant Blank. WHO CREATES CONTENT?: Stratification and content creation on the Internet. *Information, Communication & Society*, 16(4):590–612, May 2013. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2013.777758. URL http://www.tandfonline.com/doi/abs/10.1080/1369118X.2013.777758.

Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890):695–700, December 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-04198-4. URL https://www.nature.com/articles/s41586-021-04198-4.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Christo Buschek and Jer Thorp. Models All The Way Down. URL https://knowingmachines.org/models-all-the-way.

Roberto Cerina and Raymond Duch. Artificially Intelligent Opinion Polling, September 2023. URL http://arxiv.org/abs/2309.06029. arXiv:2309.06029 [stat].

Line H. Clemmensen and Rune D. Kjærsgaard. Data Representativity for Machine Learning and AI Systems, February 2023. URL http://arxiv.org/abs/2203.04706. arXiv:2203.04706 [cs, stat].

Frederick G. Conrad, Michael F. Schober, Matt Jans, Rachel A. Orlowski, Daniel Nielsen, and Rachel Levenstein. Comprehension and engagement in survey interviews with virtual agents. *Frontiers in Psychology*, 6, October 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.01578. URL http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.01578/abstract.

Frederick G. Conrad, Michael Schober, Daniel Nielsen, and Heidi Reichert. Race-of-Virtual-Interviewer Effects. 2019. URL https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1011&context=sociw.

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

Molly Crockett and Lisa Messeri. Should large language models replace human participants?, June 2023. URL https://osf.io/4zdx9.

Alejandro Cuevas, Eva M. Brown, Jennifer V. Scurrell, Jason Entenmann, and Madeleine I. G. Daepp. Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models, October 2023. URL http://arxiv.org/abs/2309.10187. arXiv:2309.10187 [cs].

Jessica Daikeler, Leon Fröhling, Indira Sen, Lukas Birkenmaier, Tobias Gummer, Jan Schwalbach, Henning Silber, Bernd Weiß, Katrin Weller, and Clemens Lechner. Assessing Data Quality in the Age of Digital Social Research: A Systematic Review. *Social Science Computer Review*, pp. 1–37, 2024. doi: 10.1177/08944393241245395. URL https://journals.sagepub.com/doi/epub/10.1177/08944393241245395.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. Using large language models in psychology. *Nature Reviews Psychology*, October 2023. ISSN 2731-0574. doi: 10.1038/s44159-023-00241-5. URL https://www.nature.com/articles/s44159-023-00241-5.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 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, December 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2309583120. URL https://pnas.org/doi/10.1073/pnas.2309583120.

Catherine D'Ignazio and Lauren F. Klein. *Data Feminism*. The MIT Press, March 2020. ISBN 978-0-262-35852-1. doi: 10.7551/mitpress/11805.001.0001. URL https://doi.org/10.7551/mitpress/11805.001.0001. _eprint: https://direct.mit.edu/book-pdf/2390355/book_9780262358521.pdf.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, July 2023. ISSN 13646613. doi: 10.1016/j.tics.2023.04.008. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661323000980.

David Dutwin and Trent D. Buskirk. A Deeper Dive into the Digital Divide: Reducing Coverage Bias in Internet Surveys. *Social Science Computer Review*, 41(5):1902–1920, October 2023. ISSN 0894-4393, 1552-8286. doi: 10.1177/08944393221093467. URL https://journals.sagepub.com/doi/10.1177/08944393221093467.

Jens Foerderer. Should we trust web-scraped data?, August 2023. URL http://arxiv.org/abs/2308.02231. arXiv:2308.02231 [cs, econ, q-fin, stat].

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. Personas with Attitudes: Controlling LLMs for Diverse Data Annotation, October 2024. URL http://arxiv.org/abs/2410.11745. arXiv:2410.11745.

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

Sourojit Ghosh and Aylin Caliskan. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, pp. 397–415, August 2023. doi: 10.1145/3568813.3600120. URL http://arxiv.org/abs/2305.10510. arXiv:2305.10510 [cs].

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, July 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2305016120. URL http://arxiv.org/abs/2303.15056. arXiv:2303.15056 [cs].

Nicole Gross. What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI. *Social Sciences*, 12(8):435, August 2023. ISSN 2076-0760. doi: 10.3390/socsci12080435. URL https://www.mdpi.com/2076-0760/12/8/435.

Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. AI and the transformation of social science research. *Science*, 380(6650):1108–1109, June 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adi1778. URL https://www.science.org/doi/10.1126/science.adi1778.

R. M. Groves and L. Lyberg. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74(5):849–879, January 2010. ISSN 0033-362X, 1537-5331. doi: 10.1093/poq/nfq065. URL https://academic.oup.com/poq/article-lookup/doi/10.1093/poq/nfq065.

Robert M. Groves, Floyd. J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Robert Tourangeau. *Survey Methodology*. John Wiley & Sons, 2009.

George Gui and Olivier Toubia. The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4650172. URL https://www.ssrn.com/abstract=4650172.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs, January 2024. URL http://arxiv.org/abs/2311.04892. arXiv:2311.04892 [cs].

Friedrich M. Götz, Rakoen Maertens, Sahil Loomba, and Sander Van Der Linden. Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, February 2023. ISSN 1939-1463, 1082-989X. doi: 10.1037/met0000540. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000540.

Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. "Fifty Shades of Bias": Normative Ratings of Gender Bias in GPT Generated English Text. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1862–1876, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.115. URL https://aclanthology.org/2023.emnlp-main.115/.

Amit Haim, Alejandro Salinas, and Julian Nyarko. What's in a Name? Auditing Large Language Models for Race and Gender Bias, February 2024. URL http://arxiv.org/abs/2402.14875. arXiv:2402.14875 [cs].

Eszter Hargittai. 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, May 2015. ISSN 0002-7162, 1552-3349. doi: 10.1177/0002716215570866. URL http://journals.sagepub.com/doi/10.1177/0002716215570866.

Eszter Hargittai. Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, 38(1):10–24, 2020. doi: 10.1177/0894439318788322. URL https://journals.sagepub.com/doi/epub/10.1177/0894439318788322.

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

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation, January 2023. URL http://arxiv.org/abs/2301.01768. arXiv:2301.01768 [cs].

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. Multilingual Language Models are not Multicultural: A Case Study in Emotion. In Jeremy Barnes, Orphée De Clercq, and Roman Klinger (eds.), *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pp. 202–214, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wassa-1.19. URL https://aclanthology.org/2023.wassa-1.19/.

Ivan Hernandez and Weiwen Nie. The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, pp. peps.12543, October 2022. ISSN 0031-5826, 1744-6570. doi: 10.1111/peps.12543. URL https://onlinelibrary.wiley.com/doi/10.1111/peps.12543.

Michael Heseltine and Bernhard Clemm von Hohenberg. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1): 20531680241236239, January 2024. ISSN 2053-1680. doi: 10.1177/20531680241236239. URL https://doi.org/10.1177/20531680241236239. Publisher: SAGE Publications Ltd.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting Results of Social Science Experiments Using Large Language Models. 2024.

Christian Pieter Hoffmann, Christoph Lutz, and Miriam Meckel. Content creation on the Internet: a social cognitive perspective on the participation divide. *Information, Communication & Society*, 18(6):696–716, June 2015. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2014.991343. URL http://www.tandfonline.com/doi/abs/10.1080/1369118X.2014.991343.

Tobias Holtdirk, Dennis Assenmacher, Arnim Bleier, and Claudia Wagner. Fine-Tuning Large Language Models to Simulate German Voting Behaviour (Working Paper), October 2024. URL https://osf.io/udz28.

Björn E. Hommel. Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. *Personality and Individual Differences*, 213:112307, October 2023. ISSN 01918869. doi: 10.1016/j.paid.2023.112307. URL https://linkinghub.elsevier.com/retrieve/pii/S0191886923002301.

John J Horton. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? Working Paper 31122, National Bureau of Economic Research, April 2023. URL http://www.nber.org/papers/w31122. Series: Working Paper Series.

Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021. ISSN 1749-818X. doi: 10.1111/lnc3.12432. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12432. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12432.

James Huckle and Sean Williams. Easy Problems that LLMs Get Wrong. In Kohei Arai (ed.), *Advances in Information and Communication*, pp. 313–332, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-84457-7.

Jan Karem Höhne, Joshua Claassen, and Ben Lasse Wolf. LLM-driven bot infiltration: Protecting web surveys through prompt injections, 2025. URL https://jkhoehne.eu/wp-content/uploads/2025/02/hoehne-et-al-2025-LLM-driven-bot-infiltration-preprint-1.pdf.

Patricia A. Iglesias, Carlos Ochoa, and Melanie Revilla. A practical guide to (successfully) collect and process images through online surveys. *Social Sciences & Humanities Open*, 9:100792, 2024. ISSN 25902911. doi: 10.1016/j.ssaho.2023.100792. URL https://linkinghub.elsevier.com/retrieve/pii/S2590291123003972.

Rune Møberg Jacobsen, Samuel Rhys Cox, Carla F. Griggio, and Niels Van Berkel. Chatbots for Data Collection in Surveys: A Comparison of Four Theory-Based Interview Probes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, Yokohama Japan, April 2025. ACM. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3714128. URL https://dl.acm.org/doi/10.1145/3706598.3714128.

Gonzalo Jaimovitch-López, Cèsar Ferri, José Hernández-Orallo, Fernando Martínez-Plumed, and María José Ramírez-Quintana. Can language models automate data wrangling? *Machine Learning*, 112(6):2053–2082, June 2023. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-022-06259-9. URL https://link.springer.com/10.1007/s10994-022-06259-9.

Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020, September 2023. ISSN 29497191. doi: 10.1016/j.nlp.2023.100020. URL https://linkinghub.elsevier.com/retrieve/pii/S2949719123000171.

Rebecca L. Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. The Ghost in the Machine has an American accent: value conflict in GPT-3, March 2022. URL http://arxiv.org/abs/2203.07785. arXiv:2203.07785 [cs].

Andreas Jungherr, Harald Schoen, Oliver Posegga, and Pascal Jürgens. 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, June 2017. ISSN 0894-4393, 1552-8286. doi: 10.1177/0894439316631043. URL http://journals.sagepub.com/doi/10.1177/0894439316631043.

Kirill Kalinin. Improving GPT Generated Synthetic Samples with Sampling-Permutation Algorithm. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4548937. URL https://www.ssrn.com/abstract=4548937.

Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*, 71(6):922–946, December 2021. ISSN 0021-9916. doi: 10.1093/joc/jqab034. URL https://doi.org/10.1093/joc/jqab034.

Junsol Kim and Byungkyu Lee. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction, November 2023. URL http://arxiv.org/abs/2305.09620. arXiv:2305.09620 [cs].

Konstantinos Konstantis, Antonios Georgas, Antonis Faras, Konstantinos Georgas, and Aristotle Tympas. Ethical considerations in working with ChatGPT on a questionnaire about the future of work with ChatGPT. *AI and Ethics*, June 2023. ISSN 2730-5953, 2730-5961. doi: 10.1007/s43681-023-00312-6. URL https://link.springer.com/10.1007/s43681-023-00312-6.

Frauke Kreuter. Modernizing Data Collection. *Journal of Official Statistics*, 41(3):863–872, September 2025. ISSN 0282-423X, 2001-7367. doi: 10.1177/0282423X251318452. URL https://journals.sagepub.com/doi/10.1177/0282423X251318452.

Jessica B. Kuntz and Elsie C. Silva. Who Authors the Internet? Analyzing Gender Diversity in ChatGPT-3 Training Data. Technical report, Pitt Cyber – Institute for Cyber Law, Policy, and Security, September 2023.

Max M. Lang and Sol Eskenazi. Telephone Surveys Meet Conversational AI: Evaluating a LLM-Based Telephone Survey System at Scale, February 2025. URL http://arxiv.org/abs/2502.20140. arXiv:2502.20140 [cs].

Antonio Laverghetta Jr. and John Licato. Generating Better Items for Cognitive Assessments Using Large Language Models. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 414–428, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.34. URL https://aclanthology.org/2023.bea-1.34/.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(14 March):1203–1205, 2014. URL https://www.science.org/doi/10.1126/science.1248506.

Benjamin Lebrun, Sharon Temtsin, Andrew Vonasch, and Christoph Bartneck. Detecting the corruption of online questionnaires by artificial intelligence. *Frontiers in Robotics and AI*, Volume 10 - 2023, 2024. ISSN 2296-9144. doi: 10.3389/frobt.2023.1277635. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2023.1277635.

Philseok Lee, Shea Fyffe, Mina Son, Zihao Jia, and Ziyu Yao. 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, February 2023. ISSN 0889-3268, 1573-353X. doi: 10.1007/s10869-022-09864-6. URL https://link.springer.com/10.1007/s10869-022-09864-6.

Joshua Lerner. The Promise & Pitfalls of AI-Augmented Survey Research, October 2024. URL https://www.norc.org/research/library/promise-pitfalls-ai-augmented-survey-research.html.

Danny D. Leybzon, Shreyas Tirumala, Nishant Jain, Summer Gillen, Michael Jackson, Cameron McPhee, and Jennifer Schmidt. AI Telephone Surveying: Automating Quantitative Data Collection with an AI Interviewer, July 2025. URL http://arxiv.org/abs/2507.17718. arXiv:2507.17718 [cs].

Bryan Li, Samar Haider, and Chris Callison-Burch. This Land is Your, My Land: Evaluating Geopolitical Bias in Language Models through Territorial Disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3855–3871, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.213. URL https://aclanthology.org/2024.naacl-long.213.

Ashley Liew and Klaus Mueller. Using Large Language Models to Generate Engaging Captions for Data Visualizations, December 2022. URL http://arxiv.org/abs/2212.14047. arXiv:2212.14047 [cs].

Mitchell Linegar, Rafal Kocielnik, and R. Michael Alvarez. Large language models and political science. *Frontiers in Political Science*, 5:1257092, October 2023. ISSN 2673-3145. doi: 10.3389/fpos.2023.1257092. URL https://www.frontiersin.org/articles/10.3389/fpos.2023.1257092/full.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, February 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00638. URL https://doi.org/10.1162/tacl_a_00638. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00638/2336043/tacl_a_00638.pdf.

Christoph Lutz. Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies*, 1(2):141–148, April 2019. ISSN 2578-1863, 2578-1863. doi: 10.1002/hbe2.140. URL https://onlinelibrary.wiley.com/doi/10.1002/hbe2.140.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. Intersectional Stereotypes in Large Language Models: Dataset and Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8589–8597, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.575. URL https://aclanthology.org/2023.findings-emnlp.575.

Antonio Maiorino, Zoe Padgett, Chun Wang, Misha Yakubovskiy, and Peng Jiang. Application and Evaluation of Large Language Models for the Generation of Survey Questions. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5244–5245, Birmingham United Kingdom, October 2023. ACM. ISBN 979-8-4007-0124-5. doi: 10.1145/3583780.3615506. URL https://dl.acm.org/doi/10.1145/3583780.3615506.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8474–8503, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.567/.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve, September 2023. URL http://arxiv.org/abs/2309.13638. arXiv:2309.13638 [cs].

Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 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), January 2024. ISSN 2053-1680, 2053-1680. doi: 10.1177/20531680241231468. URL http://journals.sagepub.com/doi/10.1177/20531680241231468.

Erica Ann Metheney and Lauren Yehle. Exploring the Potential Role of Generative AI in the TRAPD Procedure for Survey Translation, 2024. URL http://arxiv.org/abs/2411.14472v1.

Jared Moore, Tanvi Deshpande, and Diyi Yang. Are Large Language Models Consistent over Value-laden Questions?, July 2024. URL http://arxiv.org/abs/2407.02996. arXiv:2407.02996 [cs].

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: measuring ChatGPT political bias. *Public Choice*, August 2023. ISSN 0048-5829, 1573-7101. doi: 10.1007/s11127-023-01097-2. URL https://link.springer.com/10.1007/s11127-023-01097-2.

Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. SocialStigmaQA: A Benchmark to Uncover Stigma Amplification in Generative Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21454–21462, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i19.30142. URL https://ojs.aaai.org/index.php/AAAI/article/view/30142.

Paweł Niszczota, Mateusz Janczak, and Michał Misiak. Large language models can replicate cross-cultural differences in personality. *Journal of Research in Personality*, 115:104584, April 2025. ISSN 00926566. doi: 10.1016/j.jrp.2025.104584. URL https://linkinghub.elsevier.com/retrieve/pii/S0092656625000169.

Francisco Olivos and Minhui Liu. ChatGPTest: Opportunities and Cautionary Tales of Utilizing AI for Questionnaire Pretesting. *Field Methods*, 0(0):1525822X241280574, 2024. doi: 10.1177/1525822X241280574. URL https://doi.org/10.1177/1525822X241280574. _eprint: https://doi.org/10.1177/1525822X241280574.

Joseph T Ornstein, Elise N Blasingame, and Jake S Truscott. How to Train Your Stochastic Parrot: Large Language Models for Political Texts, 2024. URL https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf.

Ruby Ostrow and Adam Lopez. LLMs Reproduce Stereotypes of Sexual and Gender Minorities, January 2025. URL http://arxiv.org/abs/2501.05926. arXiv:2501.05926 [cs] version: 1.

M. A. Palacios Barea, D. Boeren, and J. F. Ferreira Goncalves. 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*, November 2023. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-023-01804-z. URL https://link.springer.com/10.1007/s00146-023-01804-z.

Beth-Ellen Pennell, Kristen Cibelli Hibben, Lars E. Lyberg, Peter Ph. Mohler, and Gelaye Worku. A Total Survey Error Perspective on Surveys in Multinational, Multiregional, and Multicultural Contexts. In Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, and Brady T. West (eds.), *Total Survey Error in Practice*, pp. 179–201. John Wiley & Sons, Inc., Hoboken, NJ, USA, February 2017. ISBN 978-1-119-04170-2 978-1-119-04167-2. doi: 10.1002/9781119041702.ch9. URL https://onlinelibrary.wiley.com/doi/10.1002/9781119041702.ch9.

Pouya Pezeshkpour and Estevam Hruschka. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. TopicGPT: A Prompt-based Topic Modeling Framework. In Kevin Duh, Helena Gomez, and Steven 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, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.164. URL https://aclanthology.org/2024.naacl-long.164/.

Yao Qu and Jue Wang. Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1095, August 2024. ISSN 2662-9992. doi: 10.1057/s41599-024-03609-x. URL https://www.nature.com/articles/s41599-024-03609-x.

Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 428–446, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.26. URL https://aclanthology.org/2023.acl-long.26/.

Luca Rettenberger, Markus Reischl, and Mark Schutera. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):42, February 2025. ISSN 2432-2725. doi: 10.1007/s42001-025-00376-w. URL https://doi.org/10.1007/s42001-025-00376-w.

Melanie Revilla, Carlos Ochoa, Jan Karem Höhne, and Mick P Couper. Transcribing and Coding Voice Answers Obtained in Web Surveys: Comparing Three Leading Automatic Speech Recognition Tools and Human versus LLM-based Coding. 2025. doi: 10.13140/RG.2.2.15968.39681. URL https://rgdoi.net/10.13140/RG.2.2.15968.39681. Publisher: Unpublished.

Gregory Roberts. AI Training Datasets: the Books1+Books2 that Big AI eats for breakfast, December 2022. URL https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/. Section: AI Reality 2022.

Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. Towards Coding Social Science Datasets with Language Models, June 2023. URL http://arxiv.org/abs/2306.02177. arXiv:2306.02177 [cs].

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

Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. Demonstrations of the Potential of AI-based Political Issue Polling. *Harvard Data Science Review*, 5(4), October 2023. Publisher: The MIT Press.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, July 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Marko Sarstedt, Susanne J. Adler, Lea Rau, and Bernd Schmitt. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6):1254–1270, 2024. ISSN 1520-6793. doi: 10.1002/mar.21982. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.21982. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21982.

Tim Schott, Daniel Furman, and Shreshta Bhat. Polyglot or Not? Measuring Multilingual Encyclopedic Knowledge in Foundation Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11238–11253, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.691. URL https://aclanthology.org/2023.emnlp-main.691.

Shannon Schumacher and Nicholas Kent. 8 charts on internet use around the world as countries grapple with COVID-19, October 2023. URL https://www.pewresearch.org/short-reads/2020/04/02/8-charts-on-internet-use-around-the-world-as-countries-grapple-with-covid-19/.

Howard Schuman and Stanley Presser. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage, 1996.

Indira Sen, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. A TOTAL ERROR FRAMEWORK FOR DIGITAL TRACES OF HUMAN BEHAVIOR ON ONLINE PLATFORMS. *Public Opinion Quarterly*, 85:399–422, 2021. doi: 10.1093/poq/nfab018. URL https://doi.org/10.1093/poq/nfab018.

Aaron Shaw and Eszter Hargittai. The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing. *Journal of Communication*, 68(1):143–168, February 2018. ISSN 0021-9916, 1460-2466. doi: 10.1093/joc/jqx003. URL https://academic.oup.com/joc/article/68/1/143/4915319.

Gabriel Simmons and Christopher Hare. Large Language Models as Subpopulation Representative Models: A Review, 2023. URL https://doi.org/10.48550/arXiv.2310.17888. https://arxiv.org/abs/2310.17888.

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

Trevon Tewari and Patrick Hosein. Automating the Conducting of Surveys Using Large Language Models. In Ana Fred, Allel Hadjali, Oleg Gusikhin, and Carlo Sansone (eds.), *Deep Learning Theory and Applications*, pp. 136–151, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-66705-3.

Arun J. Thirunavukarasu and Jessica O'Logbon. The potential and perils of generative artificial intelligence in psychiatry and psychology. *Nature Mental Health*, 2(7):745–746, May 2024. ISSN 2731-6076. doi: 10.1038/s44220-024-00257-7. URL https://www.nature.com/articles/s44220-024-00257-7.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, September 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00685. URL https://doi.org/10.1162/tacl_a_00685. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00685/2468689/tacl_a_00685.pdf.

Joshua Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3144139. URL https://www.ssrn.com/abstract=3144139.

Petter Törnberg. Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. *Social Science Computer Review*, pp. 08944393241286471, September 2024. ISSN 0894-4393. doi: 10.1177/08944393241286471. URL https://doi.org/10.1177/08944393241286471. Publisher: SAGE Publications Inc.

International Telecommunication Union. Measuring digital development Facts and Figures 2022. Technical report, 2022.

Yamil Ricardo Velez. Crowdsourced Adaptive Surveys. *Political Analysis*, pp. 1–14, 2025. doi: 10.1017/pan.2024.34. URL https://www.cambridge.org/core/journals/political-analysis/article/crowdsourced-adaptive-surveys/EC492E9F650ADB1A24CBF49F93540436.

Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David Rothschild, and Robert West. Prevalence and Prevention of Large Language Model Use in Crowd Work. *Commun. ACM*, 68(3):42–47, February 2025. ISSN 0001-0782. doi: 10.1145/3685527. URL https://doi.org/10.1145/3685527. Place: New York, NY, USA Publisher: Association for Computing Machinery.

Leah von der Heyde, Anna-Carolina Haensch, Alexander Wenz, and Bolei Ma. United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections, 2024. URL https://arxiv.org/abs/2409.09045. Version Number: 2.

Leah von der Heyde, Anna-Carolina Haensch, Bernd Weiß, and Jessica Daikeler. AIn't Nothing But a Survey? Using Large Language Models for Coding German Open-Ended Survey Responses on Survey Motivation, 2025. URL https://arxiv.org/abs/2506.14634. Version Number: 2.

Christina P Walker and Joan C Timoneda. Identifying the sources of ideological bias in GPT models through linguistic variation in output.

Chenglong Wang, Bongshin Lee, Steven Drucker, Dan Marshall, and Jianfeng Gao. Data Formulator 2: Iterative Creation of Data Visualizations, with AI Transforming Data Along the Way, February 2025. URL http://arxiv.org/abs/2408.16119. arXiv:2408.16119 [cs].

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6349–6384, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.345. URL https://aclanthology.org/2024.acl-long.345/.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

Brock Webb. Synthetic Survey Respondents Creator, August 2024. URL https://github.com/brockwebb/Synthetic-Survey-Respondents-Creator.

Alexander Wuttke, Matthias Aßenmacher, Christopher Klamm, Max M. Lang, Quirin Würschinger, and Frauke Kreuter. AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers, September 2024. URL http://arxiv.org/abs/2410.01824. arXiv:2410.01824 [cs].

Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Transactions on Computer-Human Interaction*, 27(3):1–37, June 2020. ISSN 1073-0516, 1557-7325. doi: 10.1145/3381804. URL http://arxiv.org/abs/1905.10700. arXiv:1905.10700 [cs].

Brahim Zarouali, Theo Araujo, Jakob Ohme, and Claes de Vreese. Comparing Chatbots and Online Surveys for (Longitudinal) Data Collection: An Investigation of Response Characteristics, Data Quality, and User Evaluation. *Communication Methods and Measures*, pp. 1–20, January 2023. ISSN 1931-2458, 1931-2466. doi: 10.1080/19312458.2022.2156489. URL https://www.tandfonline.com/doi/full/10.1080/19312458.2022.2156489.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7915–7927, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.491. URL https://aclanthology.org/2023.emnlp-main.491/.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When "A Helpful Assistant" Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15126–15154, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.888. URL https://aclanthology.org/2024.findings-emnlp.888/.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291, March 2024. doi: 10.1162/coli_a_00502. URL https://aclanthology.org/2024.cl-1.8/. Place: Cambridge, MA Publisher: MIT Press.

Zhao Zou, Omar Mubin, Fady Alnajjar, and Luqman Ali. A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires. *Scientific Reports*, 14(1):2781, February 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-53255-1. URL https://www.nature.com/articles/s41598-024-53255-1.

Ibrahim Tolga Öztürk, Rostislav Nedelchev, Christian Heumann, Esteban Garces Arias, Marius Roger, Bernd Bischl, and Matthias Aßenmacher. How Different is Stereotypical Bias Across Languages? In Rosa Meo and Fabrizio Silvestri (eds.), *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 209–229, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-74630-7.

A Appendix

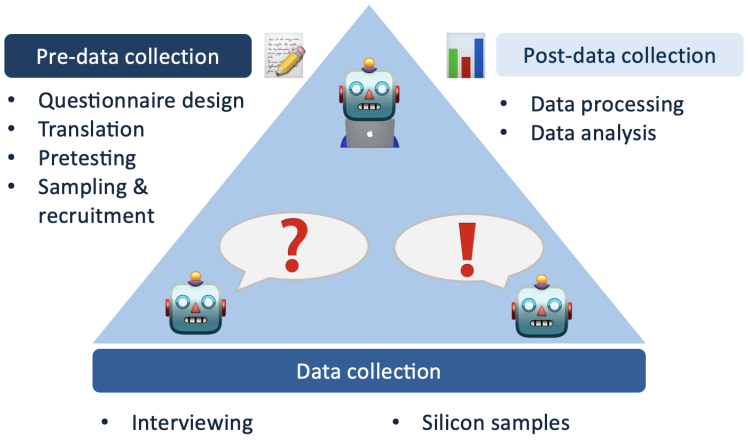

Figure A1: Roles and applications of LLMs in the survey research process.

Stage	LLM Role	Applications	Potentials	Pitfalls
Overall			• Speed • Labor-/ cost-efficiency	• Accuracy • Computational resources & expertise
Pre-data collection	**Research Assistant**	• Questionnaire design • **Questionnaire translation** • **Questionnaire evaluation** • Sampling & **recruitment material design**	• Increased scope for target-group adaptation	• Cultural, demographic, and linguistic biases
Data collection	**Interviewer**	• Augmented: Dynamic interview adaptation • **Independent: Multimodal; in-depth interviewing**	• **Flexibility (mode, content, time)** • Timeliness (emerging topics) • Standardization (emerging topics) • Accessibility (multi-mode) • **Scalability (in-depth interviews)** • Reduced interviewer effects	• Lack of rapport / human touch • Reliability
	Respondent	• **Synthetic samples**	• **Scalability** • Avoidance of respondent burden & bias (duration, sensitive topics) • Simulating hard-to-survey populations • **Timeliness (emerging topics/prediction)** • **Imputation**	• **Cultural, demographic, political, linguistic, temporal biases** • **Validity** • **Reliability** • (Lack of) variability • Non-human response behavior • Machine desirability
Post-data collection	**Research Assistant**	• **Multimodal data processing:** structuring, classification, curation • Data analysis	• **Automation** • **Scalability** • Reduced human errors	• **Cultural, demographic, and linguistic biases** • **Misinterpretation** • Reliability

Table A1: Summary of data quality potentials and pitfalls of using LLMs in survey research. *Items with major impact are boldened. LLM training data, alignment, model architecture, and research design can affect data quality overall.*