Natural Language Processing Journal 4 (2023) 100020

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Employing large language models in survey research

Bernard J. Jansen [a,*], Soon-gyo Jung [a], Joni Salminen [b]

[a] Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar
[b] School of Marketing and Communication, University of Vaasa, Finland

ARTICLE INFO

Keywords:
Survey research
Large language models
Survey data
Surveys
LLM survey respondents

ABSTRACT

This article discusses the promising potential of employing large language models (LLMs) for survey research, including generating responses to survey items. LLMs can address some of the challenges associated with survey research regarding question-wording and response bias. They can address issues relating to a lack of clarity and understanding but cannot yet correct for sampling or nonresponse bias challenges. While LLMs can assist with some of the challenges with survey research, at present, LLMs need to be used in conjunction with other methods and approaches. With thoughtful and nuanced approaches to development, LLMs can be used responsibly and beneficially while minimizing the associated risks.

1. Introduction

On 31 May 2023, the company CloudResearch, a survey participant recruitment company, sent out via its company listserv the email message shown in Fig. 1.

The email message claimed that CloudResearch had addressed several persistent problems in survey research by engineering billions of simulated but unique human personalities available for behavioral research. No need for humans! CloudResearch's Chief Technology Officer Jonathan Robinson stated, "Our team has been working on this advancement for years. Survey researchers kept telling us about problems they were having with attention and data quality. It's also always been difficult to find people from hard-to-reach groups. So, we thought, 'What if we just got rid of the people altogether? That would solve a lot of problems'." (Moss, 2023). CloudResearch claimed several benefits on its blog from leveraging AI for the creation of survey participants, including (presented in a list format that looks like ChatGPT wrote it) an amazingly low 0.8% margin of error, immediate access, cost savings, superior data quality, perfect results, and expanded reach (Moss, 2023).

Although the email message and blog posting was an April Fools' Day joke, the reaction to the email message and blog posting from an informal focus group was "Oh, this is totally possible!" highlighting the potential future (near term) impact of large language models (LLM) on the domain of survey research, which is the topical impact that we discuss in this communication paper.

The debut of ChatGPT and other large language and Generative Pre-trained Transformer (GPT) models has generated significant attention from the natural language processing (NLP) community and nearly every domain that deals with words. These NLP models are trained on massive amounts of text data and can generate human-like text, answer questions, and even engage in conversations. Open AI's ChatGPT, in particular, has been hailed as a breakthrough in NLP, as it has achieved state-of-the-art performance on a wide range of language tasks. This paper will explore some of the potential benefits, drawbacks, and ethical considerations associated with using ChatGPT and other LLMs within particular and vital domains such as survey research. As survey research is one of the most common tools social scientists deploy, the potential ramifications of LLMs could be tremendous. In fact, these ramifications are worth any number of analyses and articles, of which the current manuscript is but one. The motivational question we address through our analysis is, *can generative AI improve survey research*?

2. Survey research: Process and challenges

Survey research is a research method that involves collecting data from a sample of individuals by using standardized questionnaires (called surveys or survey instruments). The goal of survey research is to gather information about the attitudes, opinions, beliefs, and behaviors of the targeted population through closed-ended questions (which result in quantitative data) and open-ended ones (which result in qualitative data) (Aldridge, 2001; Braun et al., 2021). Research using surveys can be conducted via telephone, by mail, online, or as in-person interviews. Online surveys are prevalent due to the ease with which they can be implemented and their low cost relative to other modes of collecting data (Jansen et al., 2007; Sue and Ritter, 2012). The data collected from surveys is analyzed using statistical techniques to either identify patterns, relationships, and trends in the data (Bryman and Cramer, 2002) or harness the rich potential of qualitative data through

* Corresponding author.
E-mail addresses: jjansen@acm.org (B.J. Jansen), sjung@hbku.edu.qa (S.-g. Jung), jonisalm@uwasa.fi (J. Salminen).

https://doi.org/10.1016/j.nlp.2023.100020
Received 30 May 2023; Received in revised form 9 June 2023; Accepted 9 June 2023

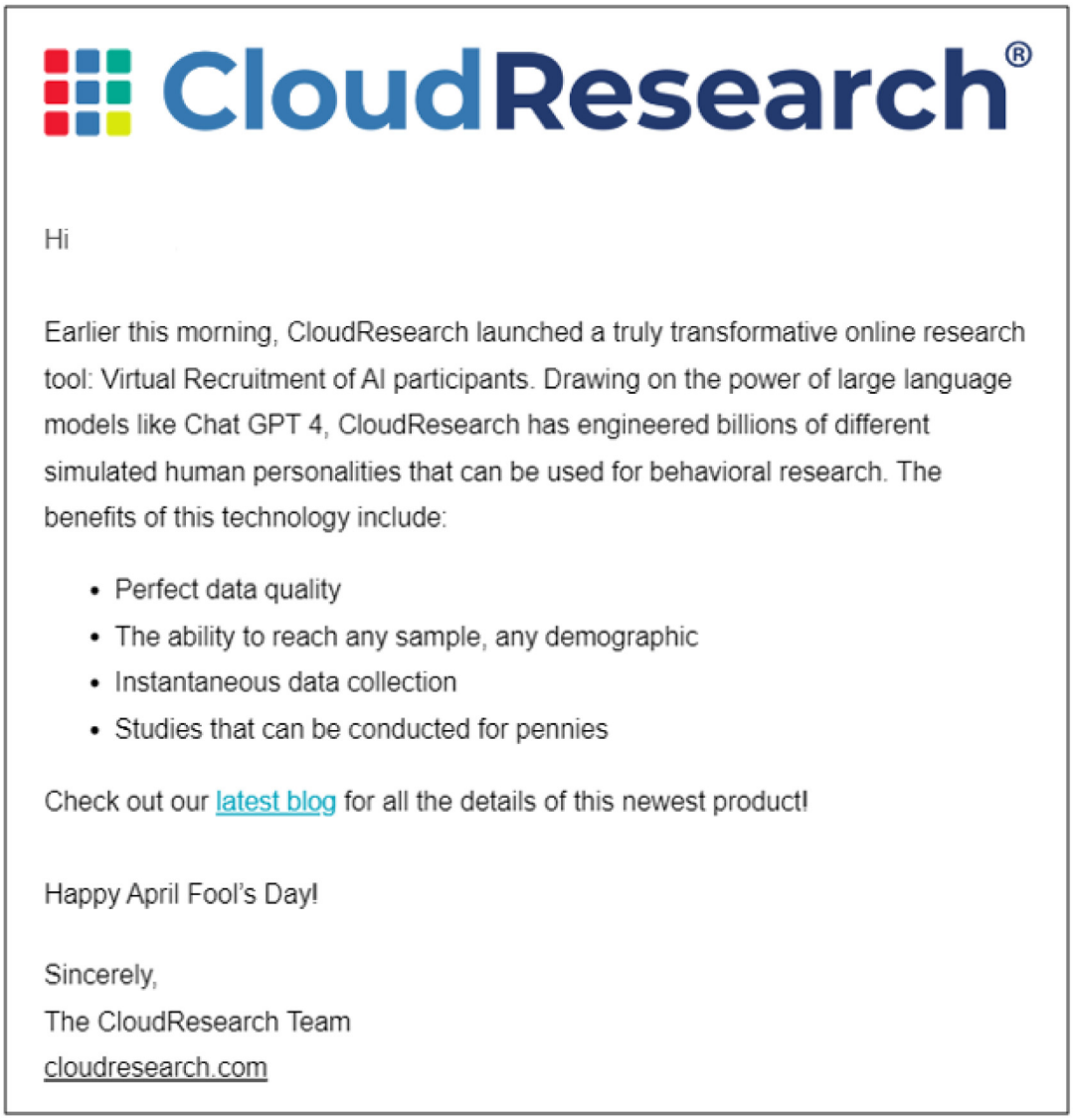

Fig. 1. Email Message from CloudResearch announcing the creation of virtual panels of survey participants (Moss, 2023).

different qualitative analyses (Braun and Clarke, 2013). Survey research is widely used in social science, marketing, information systems, human–computer interaction, and other fields where data on human attitudes and behaviors is needed. Survey analysis and reporting are increasingly leveraging machine learning (ML) for research purposes, as shown in Fig. 2, a dashboard from *Survey2Persona* (Salminen et al., 2022a), an ML learning survey analysis and visualization system.

3. Survey research and LLMs

Since survey research deals typically with words in the questions, words in the responses, or both, it is natural that LLM would impact the survey research domain. Several common tasks involved in survey research could be completed through the use of these LLM models.

- For example, *designing the survey instrument* involves developing the survey questions, response options, item construct (Salminen et al., 2020), and any other necessary components of the survey instrument — LLMs could help phrase the questions and pinpoint any inconsistencies, and perhaps suggest the best response options to measure respondents' opinions.
- *Sampling* means selecting a representative sample of individuals from the target population, which can vary depending on the research question and resources available — LLMs could suggest appropriate samples and techniques for recruiting participants. As part of *Sampling*, LLMs can perform intelligent interviewing through conversational AI instead of the conventional survey where text is read and responded to by the respondents.
- *Data cleaning and management* is processing and organizing the collected survey data to ensure its accuracy, completeness, and consistency — LLMs could, perhaps, detect inconsistent and uniform selections, resulting in low-quality entries by analyzing close-ended responses and identifying gibberish and spelling mistakes in open-ended responses.
- *Data analysis* uses statistical and qualitative methods to analyze the survey data and identify patterns, relationships, and trends in the data — there are already social media posts circulating about people using ChatGPT's Code Interpreter plugin to automate data analysis (Feng et al., 2023).
- *Reporting and dissemination* summarize the survey findings and present them in a format accessible to the target audience, such as summaries, visualizations, presentations, and even written reports — again, LLMs that can implement data science code could help facilitate this process.

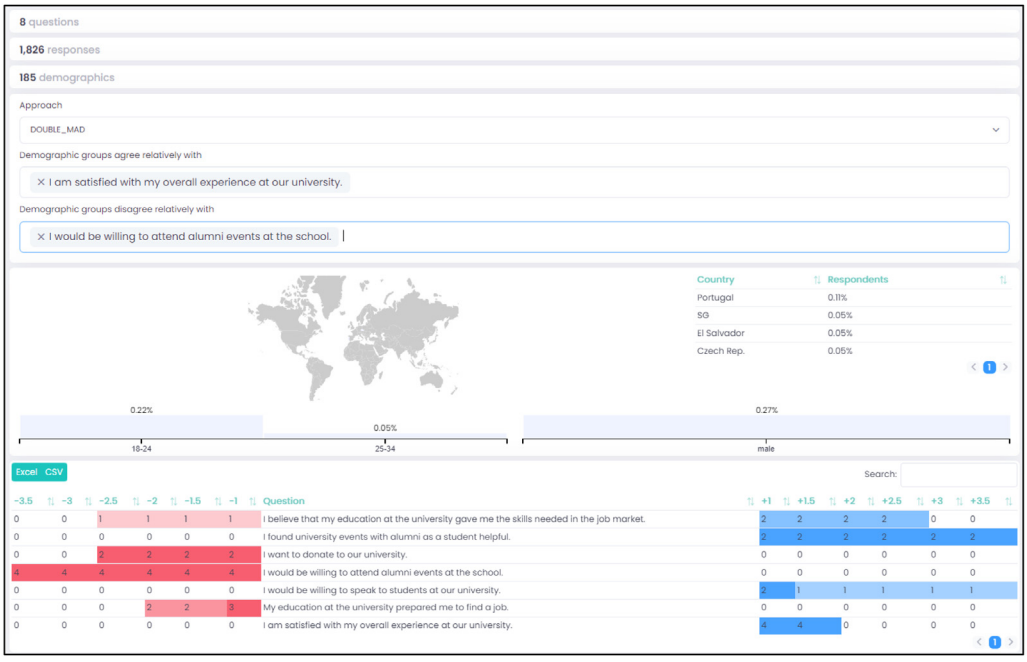

Fig. 2. ML analysis of survey data from Survey2Person (Salminen et al., 2022a).

One can easily see LLM assisting in all tasks; at least, that is the general direction in which this technology is going. Overall, survey research involves a range of language and analysis tasks that are near tailor-made for using LLMs. Using these models could potentially significantly improve the efficiency of executing these tasks. Some possible ways that LLMs could process survey responses would be simulating human responses and predicting public opinion, augmenting surveys with generative AI to create new survey questions, filling in missing data, providing feedback to respondents, and reporting survey responses as interaction data (i.e., using LLMs to capture and transmit the text of the survey questions and the responses). It remains to be seen if these models can improve the effectiveness of the execution of these tasks, as these tasks require careful planning and execution to avoid bias and ensure the accuracy and reliability of survey findings. Seemingly, the only primary survey research task these LLMs cannot yet do is data collection, which is administering the survey instrument to the selected sample. However, as our CloudResearch April Fools' Day spoof hints at, creating AI-generated responses via AI-generated simulated humans may not be far off, assuming it is not already occurring.

There are also several challenges associated with survey research that AI models can address, which would result in increased effectiveness of survey research. For example, a common issue in survey research is a lack of clarity and understanding that occurs when individuals do not fully understand the survey questions or response options, leading to inaccurate or incomplete responses. Data management and analysis challenges related to data cleaning, organization, and analysis can lead to errors or inaccuracies in the results. There are also ethical considerations related to informed consent, confidentiality, and privacy of survey respondents (Spaeth, 1992). These challenges can impact the validity and reliability of survey research findings, highlighting the importance of careful planning, execution, and analysis of survey research to minimize potential biases and ensure the accuracy of the results. Again, one can envision LLMs assisting with most, if not all, of these challenges.

4. Motivation for using LLMs in survey research

The development of LLMs (Chen et al., 2022) has the potential to revolutionize the field of survey research and bring us closer to achieving more accurate, explainable (Cambria et al., 2023), and reliable survey findings, and more efficient surveys. These models may also be able to improve NLP survey tasks and develop machines that can truly understand human language and responses to survey collection.

Gilardi et al. (2023) present evidence that ChatGPT is a suitable replacement for human annotators for various NLP annotation tasks. Their results indicated that the model's zero-shot accuracy exceeds that of crowd-workers in four out of five tasks, and its intercoder agreement was higher than that of both crowd-workers and trained annotators. Furthermore, the per-annotation cost was only $0.003, a savings of twenty times compared to Amazon MTurk (the leading crowdsourcing platform for surveys). These results highlight ChatGPT's potential to significantly reduce the amount of labor and time spent on survey research.

A study by Törnberg (2023) examined the accuracy, reliability, and bias of ChatGPT when classifying Twitter users' political affiliation based on the content of a tweet. ChatGPT was compared to annotation provided by expert classifiers and crowdsourced workers, traditionally seen as the gold standard for similar tasks. Tweets from United States politicians during the 2020 election were used as the ground truth to measure the accuracy of the LLM. The results indicated that ChatGPT outperformed human classifiers regarding accuracy and reliability and had an equal or lower bias. Crucially, the LLM could correctly analyze messages that require reasoning and interpretation based on contextual knowledge, abilities that are often seen as exclusive to humans. These findings suggest that LLMs have substantial potential for use in the social sciences, enabling interpretive research on a much larger scale.

Cegin et al. (2023) studied whether ChatGPT could potentially substitute human workers in paraphrase generation for intent classification. For this, they quasi-replicated the data collection methodology of an existing crowdsourcing study on a similar scale, prompting with the same seed data and using ChatGPT instead of human labor. The results showed that ChatGPT-created paraphrases were more diverse and could thus lead to more robust machine-learning models.

On the other hand, Bisbee et al. (2023) investigated the use of ChatGPT for measuring public opinion, showing that it is not a reliable substitute for human respondents. They found that ChatGPT-generated responses overly exaggerate the extremity and certainty of partisan and social division compared to actual opinions of those possessing the same attributes. Measurements of partisan and racial affective

polarization produced by prompted "persona" profiles in ChatGPT are seven times larger than the average human opinion, while the standard deviation of synthetic data was only 31% of the variation found among real human opinions. As these models are proprietary, the researchers could not identify the cause of the bias, but their findings raise questions about the viability of using closed-source LLMs as synthetic data.

Hämäläinen et al. (2023) explored using LLMs for generating synthetic user research data. They used the GPT-3 model to generate responses to open-ended questions on the topic of video games as art. Results showed that GPT-3 could generate plausible accounts of HCI experiences. The researchers argue that LLM-generated data can be useful in designing and assessing experiments because it is a cheap and rapid process. However, they also cautioned to double-check the correctness of any resulting conclusions with real data. Their findings also present potential concerns since LLMs could be used to use crowdsourcing services. If this were to occur, the crowdsourcing of self-reported data could become subject to unreliability.

Kim and Lee (2023) analyzed how LLMs could augment surveys and enable missing data imputation, retrodiction, and zero-shot prediction. They proposed a novel methodological framework integrating survey questions, individual beliefs, and temporal contexts to tailor LLMs for opinion prediction. Results suggested that the best models were highly accurate for missing data imputation and retrodiction. They could, for instance, help identify shifts in public support for same-sex marriage. However, the models demonstrated limited performance for zero-shot prediction. The researchers also found that accuracy was lower for people with lower socioeconomic status, non-partisan affiliations, and racial minorities yet was slightly higher for ideologically sorted opinions in contemporary periods. Thus, their results implied a need for adequate socio-demographic representation and ethical considerations related to LLM deployment.

5. Considerations of employing LLMs

Therefore, as with any new technology, there are potential benefits and drawbacks to consider. First, LLMs may be able to generate compelling fake text and findings from survey data, which could have significant implications for issues like disinformation and misinformation. LLMs' ability to generate persuasive fake text or fake results from data analysis, which could result from intentional or unintentional prompts from survey researchers when leveraging these models to summarize (Xie et al., 2023) and analyze survey results, is a critical issue. It has significant implications for research findings, including disinformation and policy implications (which often rely on survey research), as malicious actors could use these models to spread false information or impersonate real people. For example, these LLMs could create highly convincing fake responses or survey analysis results that could be erroneous, spreading misinformation. Notably, injecting artificial information into decision processes via public policy survey research remains a top risk. The issue has political dimensions, as government-funded troll factories already weaponize coordinated fake news campaigns to offset the legitimacy of institutions (Bahrini et al., 2023).

Second, there is a risk that LLMs could be used to create highly realistic fake text that could be used to harm individuals or groups, such as by spreading hate speech or inciting violence. Third, there are privacy concerns about these models, as they may be trained on sensitive or personal survey data that could be used to identify individuals. Fourth, there are serious data concerns that actual (human) survey respondents would not answer the survey questions themselves but instead rely on models like ChatGPT to provide question answers using the survey items as prompts. In this scenario, the survey would not actually be the survey participant's responses, but the researcher would have no reasonable way of determining this deception.

As a result, it is important to carefully consider the ethical implications of using LLMs and to ensure that they are used responsibly and beneficially. The potential benefit (and threat) of LLMs like ChatGPT is their ability to generate highly realistic and human-like text. This capability can be employed in surveys for crafting the survey items or summarizing survey results from the analysis of survey data, all of which LLMs can do. This could have significant implications for the survey research field as machines can generate indistinguishable survey items from those written by humans. The threat is that this could significantly affect the survey research field. Relatedly, there are also concerns about the potential of these models to perpetuate biases in language data (Chakravarthi et al., 2023). For example, if an LLM is trained on text biased against certain groups of people (Diaz et al., 2018), it may reproduce those biases in its output when generating survey questions or responses. As a result, it is important for survey researchers to carefully consider the data used to train these models and ensure that they are not reinforcing harmful stereotypes or biases — predominantly, one needs to remain *critical* about the LLM outputs and not get complacent about them. Although, this understanding may be beyond the capabilities of those employing these models, as determining the biases of the outputs in real time is not a straightforward feat.

One potential way to address the issue of bias in LLMs is through the use of diverse and representative training data by those training these models. Incorporating a wide range of perspectives and voices in the training data may help minimize the risk of perpetuating harmful biases. Additionally, AI researchers can use techniques like debiasing algorithms and adversarial training to mitigate the effects of bias in language data. Another potential solution is to involve diverse experts and stakeholders in developing and evaluating LLMs, including individuals from underrepresented communities and those directly impacted by these models. Finally, survey researchers can ensure that they are not reinforcing harmful stereotypes or biases by carefully reviewing survey items through a diverse group of (human) survey researchers and editing the LLM text, which might be the most fruitful approach.

Overall, LLMs have the potential to significantly improve NLP tasks of survey-based research, such as machine translation, sentiment analysis of responses, topical classification of responses, summarization of open-ended question responses, and question composition of the survey items themselves. By training on massive amounts of text data, these models can learn to recognize complex patterns and relationships in language that may not be immediately apparent to humans (Yang et al., 2023). Additionally, the ability of these models to generate highly realistic and human-like text could have significant implications for fields like survey research, both positive and negative. For example, many survey researchers rely on participant recruitment companies with panels of participants who sign up to conduct surveys for a monetary reward (Salminen et al., 2022b). These panelists could easily leverage models like ChatGPT to respond to surveys. The result is that the data from these surveys would not be the true responses of the participants themselves. In this scenario, survey participants could submit AI-generated responses, with survey researchers using AI to analyze the responses.

Regardless, it is a scenario that survey researchers will increasingly have to face, and we expect this is already occurring in survey research as of this manuscript's preparation date. As such, using LLMs to generate survey responses deserves additional consideration.

6. Advantages of employing LLMs for survey responses

There are potential advantages to using LLMs like ChatGPT for survey research to generate survey responses. The scalability of LLMs is impressive, with these models able to generate responses to survey questions quickly and at a large scale, which can be useful for conducting surveys with many participants or generating responses to open-ended survey questions. The models are also fairly consistent (Gilardi et al., 2023); unlike human respondents, LLMs can provide consistent responses to survey questions, which can be particularly

useful for standardizing responses and minimizing variation between responses. Indeed, this is a cost-effective approach, as it eliminates the need for hiring and compensating human survey respondents. This incentivizes researchers to use LLMs, as cost is a constant issue in survey research (Salminen et al., 2022b). Also, LLMs are quite flexible in generating responses to survey questions in multiple languages, making them helpful in conducting surveys in multilingual contexts or with participants who speak different languages. These models might also be able to provide insights into language patterns and trends that may be difficult to identify through human survey responses, such as changes in word usage over time or the emergence of new language conventions. So, while there are potential challenges and limitations associated with using LLMs for survey research, these NLP models also offer several advantages that may make them valuable in the survey researcher's toolkit.

7. Potential issues of employing LLMs in survey responses

Of course, potential issues arise from using LLMs in survey research to generate survey responses. There may be bias in the language models as LLMs are trained on massive amounts of text data, which can amplify biases present in the training data. This can result in biased language generation, social stereotyping, unfair discrimination, and exclusionary norms, and it may skew survey research results (Weidinger et al., 2022). LLMs may also suffer from a lack of contextual understanding and common sense reasoning abilities. This shortcoming can generate nonsensical or inappropriate responses to survey questions (also known as 'hallucinations'). While LLMs have access to a vast generic vocabulary, they often have a limited vocabulary within a specific vertical. These models may still struggle with rare or domain-specific terms (Morozovskii and Ramanna, 2023) that may be common in survey research. This can result in inaccurate or incomplete responses to survey questions. However, a perhaps even more dangerous situation is the case of "compelling misinformation" (Spitale et al., 2023), referring to situations where the LLM produces highly convincing text that is factually wrong. Spitale et al. (2023) tested whether people can determine whether a tweet is organic and written by a Twitter user or synthetic and generated by GPT-3. The results showed that GPT-3 is capable of both creating accurate information that is clearer to understand as well as more convincing disinformation. Furthermore, people could not tell the difference between tweets generated by GPT-3 and those written by humans. So, unless the source of information divulges that it was wholly or partially generated using an LLM, people might have no way of knowing.

Apart from the above, the lack of transparency of LLMs is another major concern, as the inner workings of LLMs are often opaque and difficult to interpret. Transparency refers to the issue mentioned above of disclosing LLM participation and the intractability of LLM training and the text-generation process, sometimes called algorithmic opacity (Eslami et al., 2019). This lack of transparency makes it challenging to identify the sources of potential errors or biases in the generated responses, and this can make it challenging to validate the results of survey research.

There are also ethical considerations, as using LLMs in survey research raises concerns about using AI-generated responses to replace human participants. For example, would using LLMs as survey respondents in psychology research be appropriate or even acceptable for the research community? Would LLMs be able to mimic human-like cognition and emotions while responding to a survey involving psychology and behavior-related research?

Overall, while LLMs have demonstrated impressive capabilities in generating human-like responses, several potential issues must be considered when using them in the context of survey research. These issues highlight the need for careful consideration of the strengths and limitations of these models, as well as the potential impact of their use in survey research and resulting implications.

8. Future of LLMs in survey research

One thing is apparent — LLMs will impact survey research, not just in response generation. These models have already impacted survey research. The authors of this paper have employed LLMs in survey research in multiple ways, such as converting survey questions to statements (e.g., "Do you like ice cream?" to "I like ice cream".) and in algorithmically-generated personas from survey data, as shown in Fig. 3.

LLMs offer several potential future directions for survey research. First, the increased use of technology in survey research will likely continue to grow. This includes using online, mobile, and other digital technologies to collect survey data. The use of artificial intelligence and machine learning algorithms may also be used to help improve the accuracy and efficiency of survey research. Second, survey researchers may begin to explore non-traditional data sources, such as social media data, web analytics, and other digital data sources, to supplement or replace traditional survey data, given that LLMs can rapidly make sense of this data. Third, survey researchers may begin to integrate data from multiple sources, such as survey data, administrative data, and other data sources, to gain a more comprehensive understanding of the research question using LLMs to aid in integrating these disparate data sources. Fourth, there may be a greater emphasis on data quality in survey research, focusing on improving data collection methods, reducing nonresponse bias, and increasing response rates, perhaps using LLMs to partially address these issues. Finally, LLMs may lead to an increased focus on collaborative research in survey research, with researchers from different disciplines working together to address complex research questions.

The implications of employing LLMs might be profound, leading to a significant advance in survey research. For example, LLMs may improve the design of survey questions and response options. For example, these models may generate more neutral and objective questions or suggest a wider range of response options less likely to influence how individuals respond. NLP techniques inherent in these models may be used to analyze and interpret survey responses, allowing researchers to gain deeper insights into the data. These techniques may include sentiment analysis, topic modeling, or other NLP techniques to identify patterns and trends in the data. LLMs may be used to personalize surveys to individual respondents, tailoring the questions and response options to their individual characteristics and preferences, and these models may be used to provide real-time feedback to survey respondents, helping to improve response rates and the accuracy of the data collected. Also, LLMs may be used to support multilingual surveys, allowing researchers to collect data from a wider range of individuals and populations.

9. Probing research question

We want to close this section with a probing question: *Can synthetic data be accurate?*

If the LLM can accurately represent people's average opinions on factors like sentiment, as some nascent work suggests (Gilardi et al., 2023), then would an LLM equally well represent the average opinions of people when polling them about any societal matter? In a sense, the LLM is *trained* on public opinion, so there is a possibility that it can reflect public opinion. Therefore, there is a possibility that the synthetic, so-called "fake" response, in fact, is correct. This possibility is often ignored by treatises that categorically reject using LLMs for public opinion studies due to the myriad of risks. While we are not arguing in favor of replacing Gallup polls with LLM polls, we do want to point out that, as researchers, we must objectively examine this new technology by analyzing the full scope of its possibilities, even those that, based on first impression, appear impossible. In theory, LLMs can represent people's opinions correctly without being a fluke (see Table 1). For example, research in controlled experiments comparing LLM and human responses could be measured in various contexts and

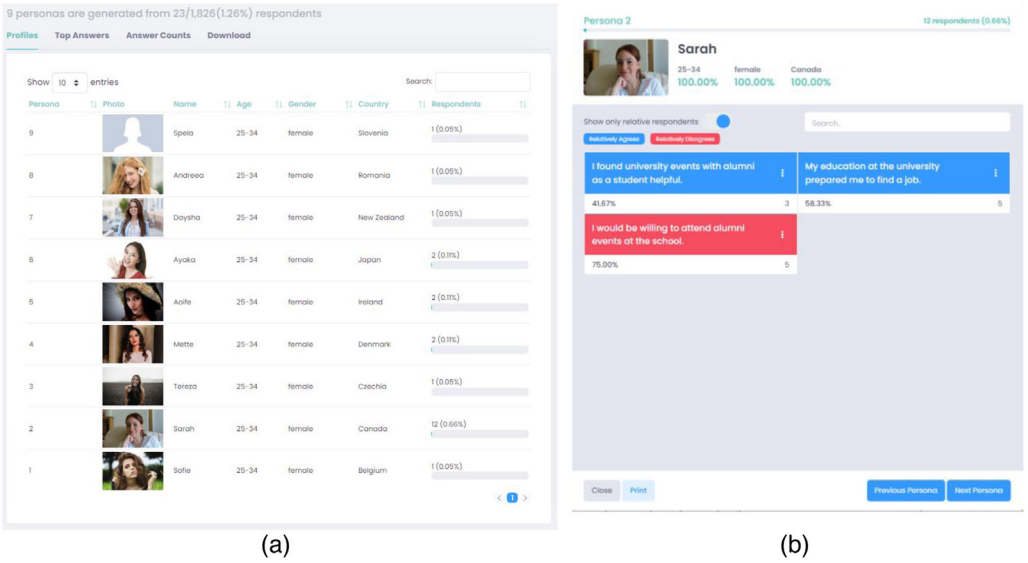

Fig. 3. Algorithmically-generated personas from Survey2Person (Salminen et al., 2022a). (a) is the personas cast (i.e., listing); (b) is a single persona profile from the cast.

Table 1
Theoretical possibilities of information accuracy by respondent source. All four options are theoretically possible. Quadrant 1 (Q1) is often refuted a priori, but we argue that more research for that quadrant is needed.

	Information given is accurate (i.e., it reflects the average opinion of the population correctly)	
Source of the information	Yes	No
LLM	Q1	Q2
Human	Q3	Q4

situations, with the accuracy of both LLM and human respondents compare to the opinions of the overall population.

As shown in Table 1, Q1 (LLM accurately reflecting the "average" human opinion of a given population) may not just be an April Fools Day joke; using LLM for survey respondents might be achievable in the near term. Q2 (LLM inaccurately reflects the "average" human opinion) and Q4 (Human respondents inaccurately reflect the "average" human opinion), if the information is inaccurate, it is not the basis for solid research. Q3 (Human respondents accurately reflecting the "average" human opinion) is, at least for now, considered the 'gold standard'. As LLM accuracy is further investigated, however, this opinion may change. An area of future research is the theoretical possibility of information accuracy, namely, how precise does a LLM have to be considered accurate?

Peer-reviewed evidence either for or against using LLMs is still too scarce to draw definitive conclusions. Our concluding statement is that LLMs will become part of the survey research process in one form or another. How extensively, we do not yet know. For now, the research community must focus on creating ethical standards and guidelines for the acceptable use of LLMs in survey research. Efforts in this area are much needed and underway (Lund et al., 2023; Pournaras, 2023; Rahimi and Abadi, 2023).

10. Conclusion

Although promising, the potential of closed-source LLMs like ChatGPT to measure human opinion has yet to be determined. While LLMs have the potential to address some of the challenges associated with survey research, they may not be a comprehensive solution to all of these challenges. They can potentially help address challenges related to question-wording and response bias by generating more neutral and objective survey questions and providing a wider range of options that are less likely to influence how individuals respond. Similarly, LLMs can potentially help address issues relating to a lack of clarity and understanding by providing more detailed explanations or examples of survey questions or response options and answering any follow-up questions that respondents might have. However, LLMs might be unable to address sampling or nonresponse bias challenges, as these issues relate more to the selection of survey respondents than to the survey questions themselves. Furthermore, ethical considerations relating to informed consent, confidentiality, and the privacy of the survey respondents are important issues that need to be carefully considered regardless of whether LLMs are used in survey research or not.

Overall, while LLMs have the potential to address some of the challenges associated with survey research, as of this writing, they should be used in conjunction with other methods and approaches (Nielsen et al., 2021; Rainie and Jansen, 2009) to ensure the accuracy and validity of the survey results. LLMs have the potential to revolutionize the field of NLP and bring us closer to developing machines that can truly understand human language. However, there are potential benefits, drawbacks, and ethical considerations associated with these models that must be carefully considered. By taking a thoughtful and nuanced approach to the development and use of LLMs, we can ensure that they are used responsibly and beneficially, maximizing their potential while minimizing the associated risks.

Acronyms

LLM: Large Language Models
NLP: Natural Language Processing
GPT: Generative Pre-trained Transformer
AI: Artificial Intelligence
ML: Machine Learning

Declaration of competing interest

No conflicts of interest.

References

Aldridge, A., 2001. Surveying the Social World: Principles and Practice in Survey Research. McGraw-Hill Education, UK.

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R.J., Esmaeili, M., Majdabadkohne, R.M., Pasehvar, M., 2023. ChatGPT: Applications, opportunities, and threats. ArXiv Preprint arXiv:2304.09103.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., Larson, J., 2023. Artificially precise extremism: How internet-trained LLMs exaggerate our differences. SocArXiv https://doi.org/10.31235/osf.io/5ecfa.

Braun, V., Clarke, V., 2013. Successful Qualitative Research: A Practical Guide for Beginners. SAGE Publications.

Braun, V., Clarke, V., Boulton, E., Davey, L., McEvoy, C., 2021. The online survey as a qualitative research tool. Int. J. Soc. Res. Methodol. Theory Pract. 24, 641–654. http://dx.doi.org/10.1080/13645579.2020.1805550.

Bryman, A., Cramer, D., 2002. Quantitative Data Analysis with SPSS Release 10 for Windows: A Guide for Social Scientists. Routledge, http://dx.doi.org/10.4324/9780203471548.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N., 2023. A survey on XAI and natural language explanations. Inf. Process. Manage. 60 (1), 103111. http://dx.doi.org/10.1016/j.ipm.2022.103111.

Cegin, J., Simko, J., Brusilovsky, P., 2023. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. ArXiv arXiv:2305.12947.

Chakravarthi, B.R., Priyadharshini, R., Banerjee, S., Jagadeeshan, M.B., Kumaresan, P.K., Ponnusamy, R., Benhur, S., McCrae, J.P., 2023. Detecting abusive comments at a fine-grained level in a low-resource language. Nat. Lang. Process. J. 3, 100006. http://dx.doi.org/10.1016/j.nlp.2023.100006.

Chen, X., Xie, H., Tao, X., 2022. Vision, status, and research topics of natural language processing. Nat. Lang. Process. J. 1, 100001. http://dx.doi.org/10.1016/j.nlp.2022.100001.

Diaz, M., Johnson, I., Lazar, A., Piper, A., Gergle, M., 2018. Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Paper 412, ACM, pp. 1–14. http://dx.doi.org/10.1145/3173574.3173986.

Eslami, M., Vaccaro, K., Lee, M.K., On, A.Elazari.Bar, Gilbert, E., Karahalios, K., 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–14.

Feng, Y., Vanam, S., Cherukupally, M., Zheng, W., Qiu, M., Chen, H., 2023. Investigating code generation performance of chat-GPT with crowdsourcing social data. In: Proceedings of the 47th IEEE Computer Software and Applications Conference. pp. 1–10.

Gilardi, F., Alizadeh, M., Kubli, M., 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. ArXiv arXiv:2303.15056.

Hämäläinen, P., Tavast, M., Kunnari, A., 2023. Evaluating large language models in generating synthetic HCI research data: A case study. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–19. http://dx.doi.org/10.1145/3544548.3580688.

Jansen, Karen J., Corley, K., Jansen, B.J., 2007. E-survey methodology. In: Handbook of Research on Electronic Surveys and Measurements. IGI Global, pp. 1–8.

Kim, J., Lee, B., 2023. AI-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys. ArXiv arXiv:2305.09620.

Lund, B., Wang, T., Mannuru, N.R., Nie, B., Shimray, S., Wang, Z., 2023. ChatGPT and a new academic reality: AI-written research papers and the ethics of the large language models in scholarly publishing. ArXiv Preprint arXiv:2303.13367.

Morozovskii, D., Ramanna, S., 2023. Rare words in text summarization. Nat. Lang. Process. J. 3, 100014. http://dx.doi.org/10.1016/j.nlp.2023.100014.

Moss, A., 2023. CloudResearch revolutionizes online survey research with virtual recruitment of AI participants. CloudResearch https://www.cloudresearch.com/resources/blog/virtual-participant-recruitment/.

Nielsen, L., Salminen, J., Jung, S.-G., Jansen, B.J., 2021. Think-aloud surveys: A method for eliciting enhanced insights during user studies. Human-Computer Interaction–INTERACT 2021, In: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part V, vol. 18, pp. 504–508.

Pournaras, E., 2023. Science in the era of ChatGPT, large language models and AI: Challenges for research ethics review and how to respond. ArXiv Preprint arXiv:2305.15299.

Rahimi, F., Abadi, A.T.B., 2023. ChatGPT and publication ethics. Arch. Med. Res. 54 (3), 272–274.

Rainie, L., Jansen, B.J., 2009. Surveys as a complementary method for web log analysis. In: HandBook of Research on Web Log Analysis. IGI Global, pp. 39–64.

Salminen, J., Jansen, J., Jung, S.-G., 2022a. Survey2Persona: Rendering Survey Responses as Personas. pp. 67–73. http://dx.doi.org/10.1145/3511047.3536403.

Salminen, J., Kamel, A.M.S., Jung, S.-G., Mustak, M., Jansen, B.J., 2022b. Fair compensation of crowdsourcing work: The problem of flat rates. Behav. Inf. Technol. 1–22.

Salminen, J., Santos, J.M., Kwak, H., An, J., Jung, S., Jansen, B.J., 2020. Persona perception scale: Development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. Int. J. Hum.-Comput. Stud. 141, 102437. http://dx.doi.org/10.1016/j.ijhcs.2020.102437.

Spaeth, J.L., 1992. Perils and Pitfalls of Survey Research (Allerton Park Institute (33rd : 1991)). Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, http://hdl.handle.net/2142/634.

Spitale, G., Biller-Andorno, N., Germani, F., 2023. AI model GPT-3 (dis)informs us better than humans. ArXiv arXiv:2301.11924.

Sue, V.M., Ritter, L.A., 2012. Conducting Online Surveys. SAGE.

Törnberg, P., 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. ArXiv arXiv:2304.06588.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., 2022. Taxonomy of risks posed by language models. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 214–229.

Xie, B., Song, J., Shao, L., Wu, S., Wei, X., Yang, B., Lin, H., Xie, J., Su, J., 2023. From statistical methods to deep learning, automatic keyphrase prediction: A survey. Inf. Process. Manage. 60 (4), 103382. http://dx.doi.org/10.1016/j.ipm.2023.103382.

Yang, Z., Liu, Y., Ouyang, C., Ren, L., Wen, W., 2023. Counterfactual can be strong in medical question and answering. Inf. Process. Manage. 60 (4), 103408. http://dx.doi.org/10.1016/j.ipm.2023.103408.