Back to the basics and to the future: Evaluating silicon samples with POR standards

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) have led to growing interest in using synthetic data for surveys. A growing body of empirical applications suggest a need to apply public opinion research (POR) best practices and standards to the evaluation of such data. To do so, we delineate synthetic data use cases by drawing parallels to survey practices. Next, we emphasize an argument-based approach to efficacy, in which a data generation process is evaluated based on specific arguments around fidelity, utility, and externality. Finally, we stress the need to critically review methodology, especially statistical conclusion validity (SCV), transparency, and reproducibility. This work-in-progress intends to facilitate conversations between computer scientists and survey practitioners by creating an evaluation framework. We intend project outputs to be a collection of open-access and living artifacts and invite others to collaborate.

1 Introduction

Generative AI (genAI) developments have led to interest in replacing or augmenting human survey responses with LLM-based "silicon samples." This project creates a framework for (a) computer scientists to understand what is expected of LLMs for survey datasets and (b) survey practitioners to evaluate synthetic survey data.

2 The rise of silicon samples and survey use cases

Between October 2022 and June 2025, researchers published over 70 empirical research and didactic pieces on using LLM-generated synthetic survey data. While some offered supportive evidence (e.g., Aher et al. 2023; Argyle et al. 2023; Dillion et al. 2023), others demonstrated concerns that such data may produce smaller variance (Bisbee et al., 2024; Dominguez-Olmedo et al., 2024; Park et al., 2024; Sun et al., 2024), mis-/under-represent certain populations (Bisbee et al., 2024; Durmus et al., 2024; Sanders et al., 2023; Santurkar et al., 2023; von der Heyde et al., 2025), reflect stereotypes (Lee et al., 2024; Santurkar et al., 2023), fail to match human mental processes (Tjuatja et al., 2024; Wang et al., 2024), or distort multivariate relationships (Bisbee et al., 2024; Dominguez-Olmedo et al., 2024; Goli & Singh, 2024; Sanders et al., 2023; von der Heyde et al., 2025).

Despite concerns, survey practitioners are increasingly asked if and how to use or create synthetic survey data. We argue applications should be organized along typical survey research use cases: (1) level-oriented population estimates (e.g., prevalence of a particular opinion), (2) structure-oriented population estimates (e.g., relationships between an opinion and a behavior), (3) estimates of between-population differences, or (4) applications that use survey data to trend or model changes or make predictions and forecasts. Guided by these specific survey use cases, evaluations can better guide practical decision making.

3 Evaluating arguments of fidelity, utility, and externality by use cases

Building on professional standards for developing and using psychological instruments (AERA/APA/NCME, 2014; Kane 2013; SIOP, 2018), we suggest an argument-based ap-

proach to evaluate synthetic survey data. First, the intended use case must be stated to define the purported interpretation or use of synthetic data. Second, the evaluation argument states the standards upon which quality is judged.

We propose three categories of standards: fidelity, utility, and externality. Fidelity considers how well synthetic survey data match the human-generated data they emulate. Common approaches to evaluate fidelity in public opinion research (POR) include comparing to gold standard benchmarks. Utility refers to synthetic survey data usefulness, given intended use AND survey cost. Survey practitioners often consider the tradeoff between cost and data quality when comparing design options. Finally, externality refers to good or bad unintended consequences from implementing a process. In POR, an example is the many free and publicly available survey datasets (e.g., American National Election Studies 2021). Considering both use cases and quality standards, we use the following proto statement as a template to construct evaluation arguments about synthetic survey data:

Proto Statement 1 (PS1): Synthetic survey data produced by {*a specific LLM-based data generation process*} is {*good / not good*} for {*a purported specific use case*} because it {*some criteria pertaining to fidelity, utility, or externality*}.

4 Methodological concerns in evaluating LLM-based synthetic data

PS1 addresses the substantive nature of an evaluation, but methodological rigor should also be considered. When evaluating synthetic survey data, concerns have been raised pertaining to statistical conclusion validity (SCV, Cook & Campbell 1979), such as the improper use of inferential statistics (Chapman, 2024). Additionally, emerging evidence of diverging "thinking processes" between silicon and human samples (Tjuatja et al., 2024; Wang et al., 2024) suggest comparison may be difficult due to a lack of measurement invariance (Meredith, 1993). Additionally, the POR community is well-aware of transparency standards (e.g., AAPOR, 2021) that require data collection processes to be documented. Transparency contributes to reproducibility, and we encourage a stronger emphasis on reproducibility. Given the constantly evolving nature of base LLMs and a lack of tractability of how these evolutions may impact synthetic data production quality, we suggest that evaluation studies incorporate planned temporal replications. These methodological considerations are represented in a second proto statement:

Proto Statement 2 (PS2): The evidence used to support Proto Statement 1 is {*sound / unsound*} because they {*meet / fail to meet*} {*some criteria pertaining to statistical conclusion validity, transparency, or reproducibility*}.

5 Framework for evaluating synthetic survey data and sharing findings

Together, PS1 and PS2 represent a proposed framework for evaluating synthetic survey data, which can be organized in a table as illustrated at `https://bit.ly/449TYTx`. Furthermore, this framework can guide the design and reporting of synthetic survey data evaluation studies by turning specific evaluation arguments into testable hypotheses and explicit quality metrics. Finally, the framework suggests that synthetic survey data evaluations are best carried out through collaboration between LLM scientists and public opinion researchers.

6 Conclusion and a call to action

We urge the NLPOR community to go "back to the basics" by grounding synthetic survey data evaluation on survey standards and human mental processes. The outputs of this project will be an open-access framework to evaluate LLM-generated synthetic survey data and a collaborative collection of evidence organized around it. We encourage others to contribute to this effort so, working together, we can provide insights on how synthetic survey data may advance survey science and business practices.

References

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. 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. 337–371. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

American Association for Public Opinion Research. American Association for Public Opinion Research Code of Professional Ethics and Practices. 2021. URL https://aapor.org/wp-content/uploads/2022/12/AAPOR-2020-Code_FINAL_APPROVED.pdf.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME). *Standards for Educational and Psychological Testing*, 2014. URL https://www.apa.org/science/programs/testing/standards.

American National Election Studies. ANES 2020 Time Series Study Full Release [dataset and documentation]. 2021. URL https://electionstudies.org/.

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*, 31(3):337–351, February 2023. ISSN 1476-4989. doi: 10.1017/pan.2023.2. URL http://dx.doi.org/10.1017/pan.2023.2.

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*, 32(4):401–416, 2024. doi: 10.1017/pan.2024.5.

Chris Chapman. "Research" concerns for LLM Applications, 2024. URL https://quantuxblog.com/research-concerns-for-llm-applications.

Thomas D. Cook and Donald Thomas Campbell. *Quasi-experimentation: Design and analysis issues for field settings*. Rand McNally, 1979.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2023.04.008. URL https://www.sciencedirect.com/science/article/pii/S1364661323000980.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 45850–45878. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/515c62809e0a29729d7eec26e2916fc0-Paper-Conference.pdf.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2024. URL https://arxiv.org/abs/2306.16388.

Ali Goli and Amandeep Singh. Can LLMs Capture Human Preferences?, 2024. URL https://arxiv.org/abs/2305.02531.

Michael T. Kane. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73, 2013. ISSN 00220655, 17453984. URL http://www.jstor.org/stable/23353796.

Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):1–14, 08 2024. doi: 10.1371/journal.pclm.0000429. URL https://doi.org/10.1371/journal.pclm.0000429.

William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993. doi: 10.1007/BF02294825.

Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770, Sep 2024. ISSN 1554-3528. doi: 10.3758/s13428-023-02307-x. URL https://doi.org/10.3758/s13428-023-02307-x.

Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. Demonstrations of the Potential of AI-based Political Issue Polling. *Harvard Data Science Review*, 5(4), oct 27 2023. https://hdsr.mitpress.mit.edu/pub/dm2hrtx0.

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, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Society for Industrial Organizational Psychology (SIOP). *Principles for the Validation and Use of Personnel Selection Procedures (5th ed.)*, 11:1–97, 12 2018. doi: 10.1017/iop.2018.195. URL https://www.apa.org/ed/accreditation/personnel-selection-procedures.pdf.

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information, 2024. URL https://arxiv.org/abs/2402.18144.

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, 2024. doi: 10.1162/tacl_a_00685. URL https://aclanthology.org/2024.tacl-1.56/.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. Vox populi, vox ai? using large language models to estimate german vote choice. *Social Science Computer Review*, 0(0):08944393251337014, 2025. doi: 10.1177/08944393251337014. URL https://doi.org/10.1177/08944393251337014.

Pengda Wang, Zilin Xiao, Hanjie Chen, and Frederick L. Oswald. Will the real linda please stand up...to large language models? examining the representativeness heuristic in LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3GhOWfSLrD.