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

PA

ARTICLE

Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson

Political Science Department, Vanderbilt University, Nashville, TN, USA

Corresponding author: James Bisbee; Email: james.h.bisbee@vanderbilt.edu

(Received 2 May 2023; revised 18 January 2024; accepted 20 January 2024; published online 17 May 2024)

Abstract

Large language models (LLMs) offer new research possibilities for social scientists, but their potential as "synthetic data" is still largely unknown. In this paper, we investigate how accurately the popular LLM ChatGPT can recover public opinion, prompting the LLM to adopt different "personas" and then provide feeling thermometer scores for 11 sociopolitical groups. The average scores generated by ChatGPT correspond closely to the averages in our baseline survey, the 2016–2020 American National Election Study (ANES). Nevertheless, sampling by ChatGPT is not reliable for statistical inference: there is less variation in responses than in the real surveys, and regression coefficients often differ significantly from equivalent estimates obtained using ANES data. We also document how the distribution of synthetic responses varies with minor changes in prompt wording, and we show how the same prompt yields significantly different results over a 3-month period. Altogether, our findings raise serious concerns about the quality, reliability, and reproducibility of synthetic survey data generated by LLMs.

Keywords: ChatGPT; synthetic data; public opinion; research ethics

Edited by: Jeff Gill