SURVEYS · TECHNOLOGY AND THE FUTURE

Synthetic Respondents and Large Language Models: Can AI Replace Survey Samples?

Some researchers and vendors now use large language models to generate answers from simulated respondents. This article explains how synthetic respondents work, what studies have found about their accuracy, why they cannot replace real samples and where they may still be useful.

ModQu research article · 2,171 words · 22 references (APA 7th edition)

Surveys are expensive and slow. Recruiting a sample, fielding a questionnaire and cleaning the data can take weeks and cost a great deal. Large language models, the AI systems behind modern chatbots, seem to offer a shortcut. If a model has learned from vast amounts of human writing, perhaps it can answer survey questions the way people would. A researcher could describe a respondent ("a 45-year-old woman in a rural area who voted for the governing party") and ask the model how that person would answer. Repeating this thousands of times with different descriptions would produce a **synthetic sample** in minutes.

Several studies and a growing number of commercial services have explored this idea. Some results are striking, others are worrying. This article explains how synthetic respondents are produced, reviews what research has found and sets out why they cannot replace real survey samples, while noting tasks where they may be useful.

How synthetic respondents are produced

A large language model is trained to predict the next word in a piece of text. During training it absorbs patterns from books, websites, news and other writing, including the ways different groups of people talk about politics, products and their lives (Bender et al., 2021).

To create a synthetic respondent, a researcher writes a **persona**, a short description of a person's characteristics, followed by a survey question. The model produces an answer that is consistent with the description. By varying the persona to match the make-up of a population, for example in age, sex, region and political views, the researcher builds a simulated sample. Answers can then be tabulated as if they came from real respondents. Figure 1 shows the steps.

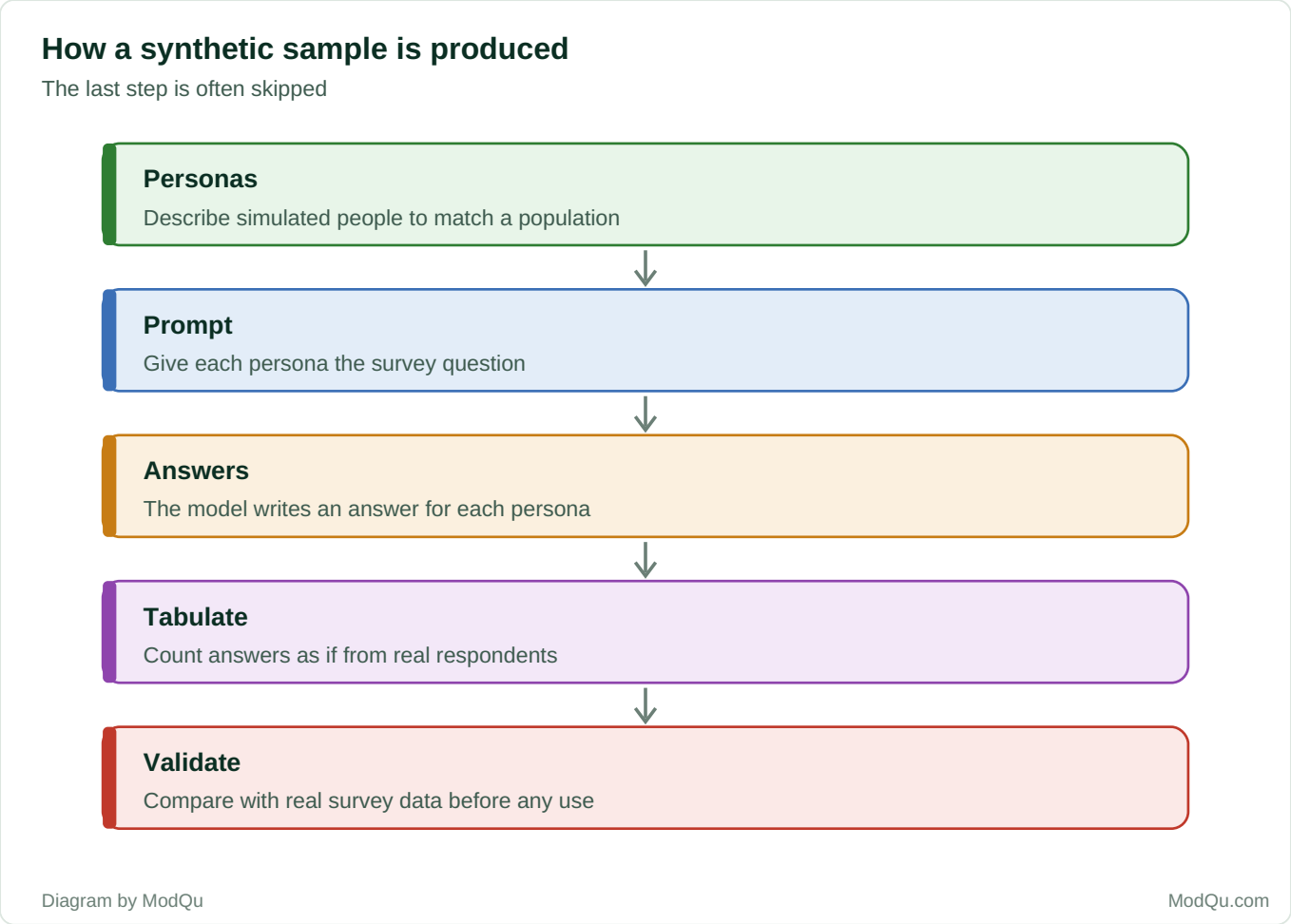


Figure 1. How a synthetic sample is produced. Diagram by ModQu.

The idea has been described in several ways. Argyle et al. (2023) called it **silicon sampling** and argued that a model's output, conditioned on a detailed persona, could reflect the views of real subgroups. Horton (2023) described language models as simulated economic agents that could be used to explore how people might respond in economic experiments. Aher et al. (2023) tested whether models could reproduce the results of classic experiments in psychology and behavioural economics. Park et al. (2023) built **generative agents**, simulated characters that remember, plan and interact over time.

Why the idea is attractive now

The interest in synthetic respondents comes at a difficult time for survey research. Response rates to many surveys have fallen for decades, costs have risen and clients expect results faster. Groves (2011) described the history of survey research as a series of eras, each shaped by the technology of its time: face-to-face interviewing, then the telephone, then the internet and new forms of data. Each change promised lower costs and faster results, and each brought new sources of error that took years to understand.

Language models are the latest technology to arrive with such promises. They are already used widely in social science for tasks such as classifying text, and a large review found that they performed well on some of these tasks and poorly on others (Ziems et al., 2024). The step from classifying what people have written to predicting what people would say is a large one, however. Classification can be checked against human judgement on the same text; a prediction about people who were never asked has nothing to be checked against unless a real survey is also carried out.

There is also a commercial pressure. A synthetic survey can be sold at a fraction of the price of a real one and delivered within hours. For decisions where a rough indication seems good enough, that offer is tempting. The question is whether the indication is accurate enough to act on, and the research reviewed below suggests that often it is not.

What studies have found

Some promising results

Argyle et al. (2023) found that, when given the demographic and political characteristics of real respondents to American election studies, a language model produced vote choices and patterns of association between variables that resembled those in the real data. They described this as "algorithmic fidelity" for some tasks. Aher et al. (2023) found that models reproduced several well-known experimental findings, such as the effect of framing on choices. Horton (2023) found that simulated agents showed patterns of behaviour similar to those found in some classic economic experiments.

These results suggest that language models capture some broad regularities in human opinion and behaviour, especially those that are widely discussed in written text.

Serious problems

Other studies have found major limitations. Bisbee et al. (2024) compared synthetic answers with answers from a large American survey on feelings towards social and political groups. Synthetic averages were sometimes close to the real ones, but the synthetic answers had far less variation than real answers, relationships between variables differed, and results changed with small changes in the wording of the prompt and between versions of the same model. They concluded that synthetic samples were not a reliable substitute for survey data.

Santurkar et al. (2023) compared the opinions expressed by several language models with those of different groups in the American population. They found that models' default opinions resembled some groups much more than others, and that asking a model to take on the views of a particular group did not fully correct this. Wang et al. (2025) found that when models were asked to represent particular identity groups, they tended to flatten the diversity within those groups and to portray them in ways that members of the groups would not recognise.

Tjuatja et al. (2024) tested whether models showed the same response biases as humans when question wording changed, such as effects of response order or of acquiescence. Models often did not respond to these changes as people do, and sometimes responded to changes that do not affect people. A synthetic sample may therefore give misleading guidance about how question design will affect real respondents.

Figure 2 illustrates a typical pattern found in comparisons: similar averages but much less spread in synthetic answers.

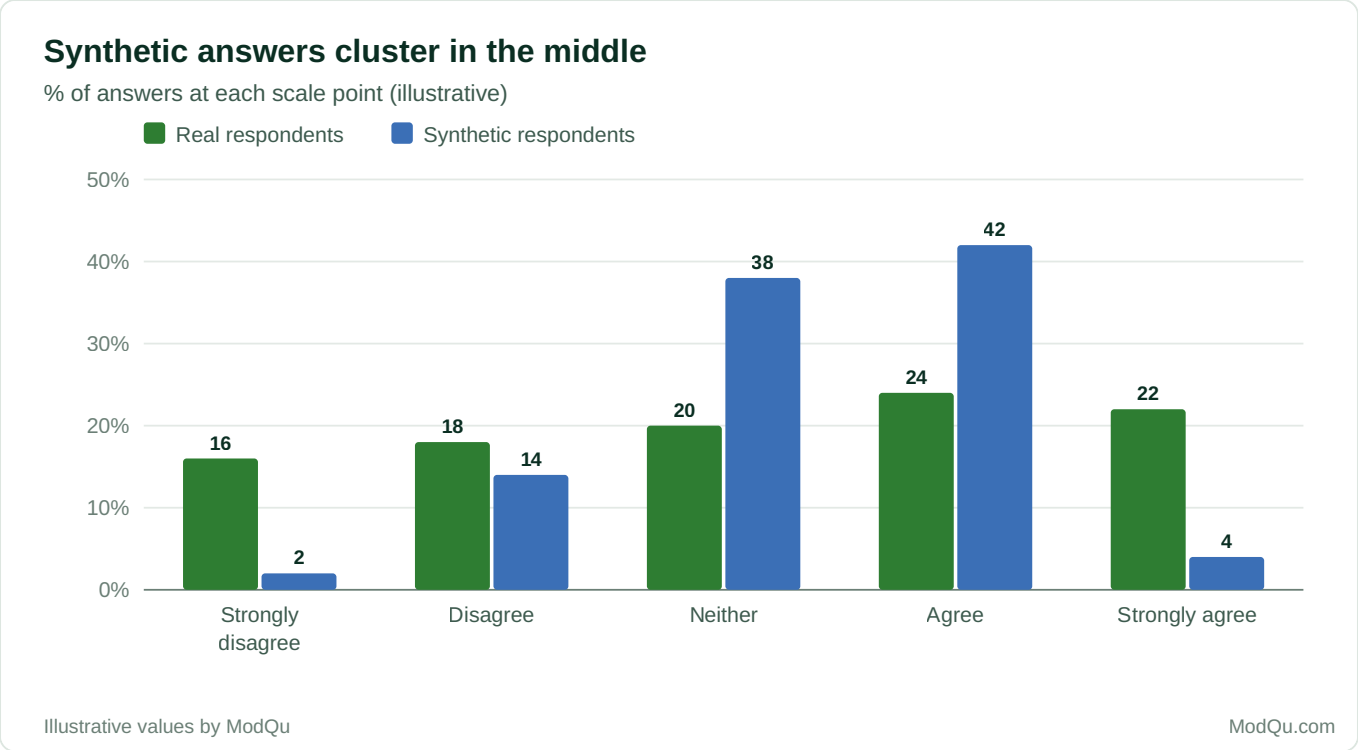


Figure 2. Distribution of answers on a five-point scale for real and synthetic respondents in a hypothetical comparison (illustrative values). Chart by ModQu.

Why synthetic respondents cannot replace samples

A model is not a population

The foundation of survey research is a sample of real people drawn from a defined population, so that what they say can be generalised to that population (Groves & Lyberg, 2010). A language model is not a sample of anything. It reflects the text on which it was trained, which over-represents people who write a great deal online, certain languages and certain countries. There is no sampling frame, no selection probability and no basis for calculating a margin of error.

Even nonprobability samples, which do not use random selection, consist of real people whose answers can be weighted and checked against known population figures (Baker et al., 2013; Mercer et al., 2017). Synthetic answers offer no such link to the population.

No new information

A survey is valuable because it measures something not already known. If a model's answers simply reflect what was in its training text, then a synthetic sample can at best repeat existing knowledge. It cannot reveal a change in opinion that happened after the model was trained, or views on a new product that people have not yet written about. Synthetic samples are weakest exactly where surveys are most needed: on new questions, emerging events and groups that are rarely heard.

Answers without a respondent

The established model of survey response describes how a person understands a question, recalls relevant information, forms a judgement and reports an answer (Tourangeau et al., 2000; Schwarz, 1999). A language model does none of this. It produces text that is statistically likely to follow the prompt. Its answers can be fluent and confident even when they are wrong, a problem known as **hallucination** (Ji et al., 2023). There is no memory of real experiences behind a synthetic answer about how often someone visited a doctor or what they paid for groceries.

Instability and opacity

Synthetic results depend on the model, its version, the wording of the persona and question, and settings that control randomness (Bisbee et al., 2024). Commercial models change without notice, so results may not be reproducible. The training data are usually not disclosed, so it is not possible to know what the model has learned about a group.

Ethical and commercial risk

Presenting synthetic answers as if they came from people would mislead clients and the public. Researchers who use synthetic data should label it clearly and never combine it with real survey data without saying so. There is also a risk that cheap synthetic results crowd out real research, especially for groups that are already under-represented. Grossmann et al. (2023) and Dillion et al. (2023) both argued that language models may be useful in social science but that their outputs must be validated against human data and must not be treated as a replacement for it.

A related problem: synthetic answers in real surveys

Language models also affect real surveys from the inside. Respondents in online panels can use AI tools to write open-ended answers, and automated programs can complete entire questionnaires with fluent, consistent text. Studies of fraudulent responses in online samples already show that bogus respondents can bias estimates (Kennedy et al., 2020), and AI makes fake answers harder to spot. Quality checks that rely on the coherence of open-ended text are becoming less reliable, so researchers need other signals, such as response timings, device checks and consistency across questions.

Where synthetic respondents may help

Although they cannot replace real samples, language models may support survey work in limited ways:

- **Early questionnaire development.** Asking a model to answer draft questions can reveal ambiguous wording or missing response options before testing with people. Such checks are a quick screen, not a substitute for cognitive interviews.
- **Generating hypotheses.** Synthetic responses may suggest possible patterns or arguments worth testing in a real survey.
- **Teaching and demonstration.** Synthetic data can illustrate analysis methods without using real personal data.
- **Testing survey software.** Simulated answers can check routing, skip logic and data processing before fieldwork.

In each case, the output is a tool for preparing real research, not a finding about real people.

Synthetic data is not the same thing

The term "synthetic data" has an older and different meaning in survey statistics. Statistical agencies create synthetic versions of real survey data to protect respondents' confidentiality. Rubin (1993) proposed replacing real records with values drawn from a statistical model fitted to the real data, so that the released file preserves the main statistical relationships while no record belongs to a real person. Similarly, **imputation** fills in missing answers using models based on the answers that were given (Little & Rubin, 2019).

These methods differ from synthetic respondents in a crucial way: they are built from, and checked against, real survey data from the population of interest. Synthetic respondents produced by a general language model have no such foundation.

How to evaluate a synthetic sample claim

Clients may be offered synthetic research by vendors. Useful questions to ask include:

- Which model and version produced the answers, and with what instructions?
- Has the method been compared with real survey data on the same topic and population, and how large were the differences in averages, spread and relationships between variables?
- How stable are the results when the prompt wording or model version changes?
- Does the topic involve events or products that arose after the model was trained?
- Are synthetic answers clearly separated from any real data in the report?

A vendor who cannot answer these questions is not offering evidence about real people.

An example

A food company wants to know how consumers would react to a new plant-based snack. A vendor offers a synthetic survey of 2,000 personas matched to the national population, delivered in a day. The company also commissions a real online survey of 1,000 consumers.

The synthetic survey finds that 58 per cent would "probably" or "definitely" try the snack, with almost no one at the extremes of the scale. The real survey finds that 41 per cent would try it, with large shares strongly for and strongly against. The real survey also reveals concern about price among younger respondents, which the synthetic personas did not mention. The company uses the synthetic results only to refine the questionnaire wording and bases its decision on the real survey (illustrative example).

Conclusion

Large language models can produce plausible answers to survey questions, and in some studies their averages resemble those of real samples. But synthetic respondents lack the essential features that make survey results trustworthy: a link to a defined population, measurement of real experience, stable and transparent methods and the ability to reveal what is

not already known. Research so far shows reduced variation, distorted subgroups and sensitivity to small changes in prompts. Synthetic respondents can help prepare research, but the generation of knowledge about people still requires asking people.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 337–371). PMLR.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology, 1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, *32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

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

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

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871. https://doi.org/10.1093/poq/nfr057

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (NBER Working Paper No. 31122). National Bureau of Economic Research. https://doi.org/10.3386/w31122

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., & Asare-Marfo, D. (2020). *Assessing the risks to online polls from bogus respondents*. Pew Research Center.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly, 81*(S1), 250–271. https://doi.org/10.1093/poq/nfw060

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (Article 2). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics, 9*(2), 461–468.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning* (pp. 29971–30004). PMLR.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93

Tjuatja, L., Chen, V., Wu, T., Talwalkar, A., & Neubig, G. (2024). Do LLMs exhibit human-like response biases? A case study in survey design. *Transactions of the Association for Computational Linguistics, 12*, 1011–1026. https://doi.org/10.1162/tacl_a_00685

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence, 7*, 400–411. https://doi.org/10.1038/s42256-025-00986-z

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics, 50*(1), 237–291. https://doi.org/10.1162/coli_a_00502