For more information contact:

Warren Chao - Chief Technology Officer

warren.chao@xpolls.ai

Can Synthetic Data Rescue Survey Research? A Case Study

Abstract

Survey research has long been a cornerstone of market insights, but data quality issues increasingly threaten its effectiveness. This paper examines the potential of synthetic data developed with AI avatars as a solution to these challenges.

This case study evaluates the reliability of AI avatars,[i] provided by Xpolls, trained with a human data set. It compares data developed from AI Avatars to actual human responses. The analysis focuses on two metrics: absolute error and matching success.

Results indicate that AI avatars precisely matched human responses 62% of the time, with a mean absolute error of 9.3 points across 25 items. Furthermore, the AI avatars matched human responses within one scale point 92% of the time, excluding seven 2-point scale items.

Those errors may seem large until human reliability is considered. The study also examined the reliability of AI avatars compared to human reliability, finding that AI avatars performed comparably to human respondents in repeated surveys.

These findings suggest that AI avatars hold promise for improving the quality and efficiency of survey research, particularly in exploratory projects and for augmenting sample sizes for hard-to-reach populations. However, further research is needed to address issues such as the AI avatars' reluctance to use the full range of response scales and to ensure the approach is suitable for multivariate analyses and academic publication.

By leveraging AI avatars, researchers may be able to overcome the challenges currently facing traditional survey research and provide a more robust and efficient method for gathering market insights in a data-centric world.

Introduction

Survey research has long been a cornerstone of market insights, but declining response rates, non-response bias, and data quality issues increasingly threaten its continued effectiveness. This paper examines the potential of AI avatars as a solution to these challenges and assesses their reliability compared to human respondents.

Situation Analysis: Data Quality in Survey Research

The field of survey research faces significant challenges, with several key issues undermining data quality. Low response rates and high non-response bias are significant concerns. Response rates have declined and are currently in the low single digits, down from 7% in 2017.[ii] The rise of bots, survey fraud, and poor-quality incentive systems further compromise data quality. Programmatic sampling and lengthy surveys contribute to these issues.[iii]

Respondents are increasingly distracted, over-surveyed, and distrustful of pollsters, with industry insiders expecting response rates to continue falling. As a result, survey research data sets can no longer rely on random or probability samples to provide projectable data. Most data sets are heavily weighted to the point where they are best described as "statistical models" and not "surveys."

These issues are causing market research clients to shift budgets towards AI-driven insights and qualitative research and away from new quantitative projects. Many end clients remain unaware of these issues or are unwilling to pay for high-quality sampling approaches. The result is an uncertain future for traditional survey research, yet the need and demand for market insights will likely increase in a data-centric world.

Situation Analysis: Artificial Intelligence

Market researchers are adopting AI to enhance efficiency and address data quality challenges. Initially, AI approaches are being used as tools to find efficiencies in the current survey research process. AI tools are helpful in questionnaire authoring and editing, automated scripting and programming, real-time probing of open-ended answers, and connecting visualization platforms for real-time data display. Large Language Models are being adapted to mine insights from existing survey research data sets.

However, these AI tools primarily focus on process improvements rather than addressing the core challenge of obtaining projectible data. Without addressing this issue, these AI tools risk developing a highly efficient garbage-in, garbage-out model, ignoring the importance of quality inputs.

Using LLM Avatars to Create Synthetic Data Sets

The use of synthetic data in survey research is not new. Researchers have been imputing missing data points for years. Choice modeling (conjoint, discrete choice) involves imputing data from limited respondent input. What is different is that in the new AI world, the scope and speed of what can be accomplished through AI has grown considerably. Large Language Models (LLMs), popularized by the launch of ChatGPT, have gained significant attention in market research.

One approach to leveraging LLMs is to develop AI avatars from a human data set. This approach starts with a data set with high response rates and low non-response bias. An "AI Avatar" is developed for each human respondent in the data set. This Avatar is essentially a "robotic twin" for each respondent. The set of AI avatars is then available to be prompted with survey questions, resulting in a data set similar in structure to human response data sets that research agencies are familiar with.

This approach has the potential to disrupt and enhance sampling in market research. It can be employed to shorten questionnaires for human respondents, ask further questions after the human data collection period, pose sensitive questions, increase sample sizes, decrease the time needed for data collection, and save budget.

Case Study Design

To evaluate the performance of AI avatars,[iv] a case study was conducted using a human data set to train AI Avatars while withholding some questions from the human data set for AI training. AI avatar performance was then compared to the corresponding human answers by posing these withheld questions to the trained AI avatars.

The human data set contained 6,002 respondents gathered in late 2022/early 2023 (HUMAN-1). This data was collected from an ABS-recruited panel recruited specifically for the client. Respondents were aged 15-24 at baseline. Question topics were wide-ranging, covering demographics, psychographics, brand awareness, past and intended behaviors, and attitudes toward substance use and other risk behaviors. Questions were posed on a variety of scales.

In addition, a second wave of human data collection occurred in fall 2023 utilizing a nearly identical questionnaire (HUMAN-2). 4,556 of the HUMAN-2 respondents were repeaters from the HUMAN-1 wave. This longitudinal data provided an opportunity to measure the AVATAR reliability against human reliability.

Working with the client, Crux Research selected 25 question items to withhold from the AI training. The goal was to remove items that crossed a range of question types, most notably past behaviors, intended future behaviors, and attitudes. The items utilized a variety of scales, from simple 2-point scales to traditional 5-point Likert scales.

In addition, we (subjectively) designated items into categories of 1) “table stakes” items, where we expected the AI avatars to match the human responses because these items were similar in content to items used to train them, 2) “probable success” items, where we felt it was reasonable to expect the AI avatars to come close, and 3) “hallucination items,” which strayed in content from the training data set and we expected the AI avatars to perform poorly.

Error Metrics

Two metrics were used to analyze the AI Avatar error: absolute error and matching success. Absolute error is a commonly used measure by academic researchers who study the accuracy of pre-election polls. It is based on the aggregated data for each question and represents how far off the AI Avatars were from the human respondent results. Matching success is the percentage of the time the AI Avatar’s response matches its human counterpart on an item-by-item level. For matching success, the percentage of the time the AI Avatar matched precisely and the percentage of the time it matched within one scale point were considered.

The case study analysis concentrated on two comparisons: the AVATAR to HUMAN-1 comparison (based on 6,002 cases) and the AVATAR error compared to the HUMAN-2 to HUMAN-1 error (based on 4,556 cases). The goal was to see how well the AI avatars matched their human counterparts and how well they matched human reliability on these items.

Results for the AVATAR to HUMAN-1 Comparison

For the AVATAR to HUMAN-1 comparison, the mean absolute error was 9.3 points across the 25 items, ranging from 0.4 to 36.4 points on individual items. The AI avatars matched the human response precisely 62% of the time. Excluding seven 2-point scale items, the AI avatars came within one scale point of the human response 92% of the time.

The errors seem due to the AI avatars' unwillingness to use the top and bottom scale points on 5-point scale items. As a result, the AI avatars’ data variability (standard deviation) was lower on our 5-point scales than that of their human counterparts’. Because of this reluctance to use the edges of the scales, the AI avatars performed better on items with fewer scale points. The avatar match rate was 48% on 5-point scales and 88% on 2-point scales.

AI avatars also did better on past behavior items (mean absolute error 6.4 points) and intended behavior items (5.6 points) than attitudinal items (14.1 points). This is likely because most attitudinal items were posed on 5-point Likert scales.

AI avatars did not display large hallucination errors. The mean absolute errors were 9.4 points on our “table stakes” items, 11.3 points on our “probable success” items, and 8.6 points on our hallucination items. The AI avatars in this study successfully answered questions that seemed beyond the scope of their training.

Subgroup Differences

Researchers, particularly those working on public health studies such as this one, should be concerned about the potential for algorithmic biases in AI approaches. An AI avatar approach must provide similar quality data across subgroups, such as race/ethnicity, LGBTQ+ status, and age.

Our findings indicate that, on average, across the 18 items we tested with more than two scale points, the AI avatar matched its human respondent within one scale point 16.2 times or 90% of the time. There was little variability in this measure across subgroups: males (16.2 matches within one scale point) and females (16.2), Whites (16.3), Blacks (16.1), Hispanics (16.0), other non-Whites (16.0), ages 15-17 (16.1), 18-20 (16.2), and 21-24 (16.2).

We considered the demographic composition of our "best" AI avatars (those that matched perfectly on 16 or more items) and our "worst" AI Avatars (those that matched perfectly on 15 or fewer items). The best AI avatars were 41% male and 59% female. The worst AI avatars were 42% male/59% female. The best AI avatars were 50% White/7% Black/16% Hispanic. The worst AI avatars were 48% White/7% Black/20% Hispanic. The age distribution was similar between the best and worst AI avatars. There was a modest difference based on education, with the Best AI avatars being slightly more educated (27% college graduates versus 21% for Worst AI avatars).

AI Avatar Reliability Compared to Human Reliability

Because this project had a second wave of human data (HUMAN-2) that surveyed many of the same people, repeated respondents can be examined to compare AI Avatar reliability with human reliability. It is perhaps unfair to expect AI avatars to match human responses when we know that human reliability in survey research is not perfect.

There were 4,556 repeated respondents with which human and AI Avatar reliability can be compared. An important limitation: Since time passed between the HUMAN-1 and HUMAN-2 data collection, some of these measures could have changed because the underlying parameters changed, and some may have more reason to change than others. We are utilizing an imperfect measure of human reliability.

For the 25 items used in this project, human reliability is just 68%; that is, repeated human respondents matched their previous answers on these 25 items about two-thirds of the time. The AI avatars matched HUMAN-1 perfectly on the 25 items 62% of the time.

Detailed results are shown in the table below.

Item Type (25 items)	% of Time Avatars Matched HUMAN-1	% of Time HUMAN-2 Matched HUMAN-1	Difference
All items	62%	68%	- 6 points
Table stakes items	59%	63%	- 4 points
Probable success items	47%	53%	- 6 points
Hallucination items	70%	75%	- 5 points
Behavioral items	75%	79%	- 4 points
Intended behavior items	68%	69%	- 1 point
Attitudinal items	46%	55%	- 9 points

When the bar is set as "within one scale point," the AI avatars compare favorably to human reliability, as shown in the table below.

Item Type (18 items not asked on 2-point scales)	% of Time Avatars Matched HUMAN-1 within 1 scale point	% of Time HUMAN-2 Matched HUMAN-1 within 1 scale point	Difference
All items	92%	90%	+ 2 points
Table stakes items	96%	93%	+ 3 points
Probable success items	91%	87%	+ 4 points
Hallucination items	94%	94%	+ 0 points
Behavioral items	95%	95%	+ 0 points
Intended behavior items	96%	95%	+ 1 point
Attitudinal items	93%	89%	+ 4 points

Takeaways and Conclusions

The absolute error of the AI avatars is about 10 points, and AI avatars match the human respondents perfectly ~62% of the time. This might seem like an unacceptable error level until human reliability is considered. Human reliability isn't perfect. In this project, humans matched themselves 68% of the time in the longitudinal study.

In broad strokes, the AI avatars were within ~6 points or so of human reliability for most items. A key lesson is that researchers need to learn how to improve question construction and prompt engineering to compel AI avatars to use the full response scales. Much of the error detected from the AI avatars was accounted for by their unwillingness to use the edges of the 5-point Likert scales.

We suspect AI avatars will soon be close to matching human reliability as researchers learn to train them and prompt them to use the full scales. This is a fast-moving field. Large Language Models continue to be refined, and survey researchers are learning how to best use them. It is fair to anticipate that the AI Avatar approach will become more feasible over time.

We suspect the demand for this approach will increase as the quality of data gathered for traditional survey research continues to decline. Many researchers question traditional data collection methods, which struggle to provide quality data and are plagued by inattentive and fraudulent respondents. The researchers' appetite for AI approaches such as this will grow.

That said, with the caveat that this will change as this approach becomes more developed, at Crux Research, we are currently making these recommendations to our clients regarding using AI avatars in survey research projects:

- **AI avatars should be considered for projects that do not require multivariate analyses.** We would not recommend using AI avatars for multivariate modeling until the issue of getting them to use the full range of larger scales is better proven, as the restricted variability this causes may not work well in multivariate modeling.

- **At the moment, we can have more confidence in behavioral and intentions questions than attitudes,** primarily because attitudinal questions tend to be measured using more scale points than behaviors. Attitudinal questions should also become more viable as we learn how to compel the AI avatars to use the edges of the scales. For now, clients should demand proof that the avatar approach will fully utilize the full range of larger scales.

- **This approach is ready for early, exploratory projects clients may conduct internally or with DIY approaches.** DIY approaches tend to be designed by less experienced researchers and often use data sources of questionable or unknown quality. Thus, they likely have higher errors than AI avatars. These projects seem to be prime candidates to replace with an AI Avatar approach.

- **This approach starts with a high-quality data set.** Synthesizing quality data requires investment in the human data set. We'd prefer to see clients invest in smaller, more tightly managed samples to seed this AI approach. This approach should only be used if clients have a starting dataset they are confident in.

- **Consider using this approach to reduce the length of questionnaires.** Shorter questionnaires allow for better-quality human data and keep respondents willing to participate in future studies. For many projects, it may be possible to gather the most critical data points via humans and then use AI avatars for less critical items.

- **Use this approach to augment your reporting.** Researchers inevitably think of questions they wish they had time to ask, or new questions arise as they analyze a human data set. This approach provides an opportunity to get those questions answered.

- **Since we found little evidence of algorithmic bias, consider using this approach to augment sample sizes for audiences that are expensive or difficult to reach:** LGBTQ+, non-whites, youth, etc.

- **Because of the cost and time savings of the avatar approach, business-to-business researchers may wish to consider it,** as incentive costs and extended data collection periods tend to limit sample sizes for these types of projects.

- **It is likely best to wait on this approach for data sets and data points destined for academic publication.** We suspect academic journals will be slow to accept synthetic data approaches.

The avatar approach holds great potential for survey research. It could become the answer researchers have been seeking to resolve the data quality problems that have emerged in our industry. A key thing to remember when using this approach is not to view it as a way to replicate the known flaws of human survey research but as a way to improve data quality. The use of AI avatars should be held to a higher standard—to obtain the true measures of human behavior and attitudes with less error than traditional survey research.

---

[i] AI Panel provided by Xpolls. Xpolls uses a proprietary method to create AI Avatars (digital twins) of real human survey respondents using respondent-specific data. Each AI Avatar is based on a single individual's opinions and attitudes derived from respondent-specific data and demographic modeling. By fine-tuning an instance of a Large Language Model to assume the identity, behavior, and beliefs of a single individual, Xpolls is able to repeat this process to reconstruct a panel of AI Avatars that correspond to an actual human panel of respondents. More information is available at Xpolls.ai.

[ii] Kennedy, Courtney, and Hannah Hartig. "Response Rates in Telephone Surveys Have Resumed Their Decline." Pew Research Center, February 27, 2019.

[iii] Deitch, JD. *The Enshittification of Programmatic Sampling*. JD Deitch, 2024.