

Leaving Insight to Digital Twins?

Promise, Progress and Limits of Synthetic Respondents

Midjourney AI prompt by GCO

AUTHORS

Carolyn Kaiser

Head of Artificial Intelligence
Nuremberg Institute for
Market Decisions (NIM)

Jakob Kaiser

Senior Researcher
Nuremberg Institute for
Market Decisions (NIM)

Rene Schallner

Founder
AI Research & Technology Lab

Vladimir Manewitsch

Senior Researcher
Nuremberg Institute for
Market Decisions (NIM)

Lea Rau

PhD Candidate and Research Assistant
Institute for Marketing
LMU Munich School of Management
Ludwig Maximilians University Munich

KEYWORDS

Synthetic Survey Data
Large Language Models
Digital Twins
Brand Funnel

Marketing teams depend on timely, reliable insights into consumer attitudes and preferences. Yet, traditional consumer surveys are becoming harder to deploy efficiently: They are costly, slow to field and increasingly affected by declining response rates. These pressures have intensified interest in scalable, low-cost alternatives that can support insight generation when time or budgets are tight. Recent advances in large language models (LLMs) have introduced one such possibility: synthetic sampling, where an LLM is asked to answer survey questions “as if” it were a consumer with a specified demographic profile. This approach creates synthetic datasets from virtual respondents – also referred to as silicon samples – and is beginning to gain attention in professional market research. If accurate, it could complement traditional surveys by enabling rapid simulations of consumer reactions or early-stage concept testing. However, existing evidence is mixed, and most validations come from non-marketing domains. An open concern is whether LLMs reflect real consumer choices and sentiment and realistic variability. We present a study with synthetic respondents, using two different measurement approaches in a realistic marketing funnel scenario (Box 1). ▶

Comparing responses from real and synthetic respondents using two different approaches

BOX 1

Study focus: consumer perceptions of soft drink brands

We developed a questionnaire focusing on consumer perceptions of soft drink brands, a category with high familiarity among U.S. consumers. The survey included four well-known brands (Coca-Cola, Pepsi, Sprite, 7UP) and four substantially less familiar brands (Dry, Moxie, Blue Sky, Orangina). This allowed us to assess whether synthetic data could appropriately differentiate responses to widely recognized versus niche brands.

Real data as a benchmark

We collected real survey data from a representative U.S. sample and surveyed 461 U.S. consumers who regularly purchased soft drinks. The sample was recruited online and quota-matched to national demographics across age, gender and ethnicity. This real-world dataset provided the baseline against which we compared the LLM-generated “digital twin” responses.

Synthetic sample: digital twins created from real demographic profiles

To generate synthetic survey responses, we created a “digital twin” for each participant that mirrored their demographic profile, including age, gender, education, occupation, financial situation, location, ethnicity and language. The LLM was then prompted to answer the same survey questions as the real participants, following the same logic as for the real consumers. This approach allowed us to produce synthetic responses that could be directly compared with real consumer data across both brand choices in the brand funnel and ratings of brand attitudes, providing a practical test of LLM-based synthetic sampling for marketing research. We used GPT-4o, the state-of-the-art model at the time of data collection, to generate the digital twins and their responses via the API with standard parameters.

Measurement: brand selection and brand attitudes

Real and synthesized participants completed a series of brand selection questions covering brand awareness, brand consideration and brand purchase. For each funnel stage, they selected specific brands. For each selected brand, human participants and digital twins answered four brand attitude questions on 7-point Likert scales. These items captured perceptions such as perceived quality and likelihood to recommend.

Taking it further: testing an alternative method – semantic similarity ratings

LLMs are optimized for generating natural language rather than discrete numerical values. Therefore, we tested whether results for Likert scales would improve when responses were first generated in text form and subsequently mapped to a numeric scale via semantic similarity rating.

How synthetic respondents performed compared to human respondents

Overall, the synthetic responses predicted individual answers significantly better than chance and successfully captured broad trends. However, synthetic data consistently overestimated positive attitudes – especially toward well-known brands – and showed significantly less variation across respondents compared with real participants.

Correct patterns, inflated likelihoods and ratings

As expected, real participants were significantly more likely to select well-known brands than lesser-known brands across the brand funnel, and the synthetic data correctly reproduced this overall pattern. However, the synthetic responses significantly overestimated brand selection for well-known brands (see Figure 1). On average, synthetic participants matched real participants’ choices

“
Synthetic data
consistently
overestimated
positive attitudes and
showed significantly
less variation across
respondents compared
with real participants.
”

The LLM generates a natural-language response to the survey question, aiming to replicate the tone and nuance of a real consumer statement. For example, when asked, “How likely is it that you would recommend this company?” the model might respond, “Soft drink brand X is a solid choice, so I’d likely recommend it.” Each Likert score is represented by a set of reference statements, also generated by the model, to capture diverse linguistic expressions for that level of intent. For instance, a score of 5 on a 7-point Likert scale might include statements such as “I’d probably recommend this company, though I’m not entirely convinced.”

To map the response to the scale, semantic similarity is computed between the generated answer and each reference statement set using text embeddings, and the Likert-scale score with the highest similarity is selected as the final rating.



Midjourney AI prompt by GCO

“
Semantic similarity
rating is an improvement,
but not a replacement
for real data.
”

about 79% of the time, showing that while LLM-generated data captures overall patterns, it tends to significantly exaggerate consumers’ likelihood of selecting brands. A similar pattern appeared for brand attitude ratings when the model provided direct Likert-scale responses. Real participants rated well-known brands higher than lesser-known ones, and synthetic data reflected this trend. Yet, synthetic ratings were systematically more positive for both brand types, with the largest bias toward well-known brands (see Figure 2). Overall, on a 7-point Likert scale, synthetic answers deviated from real answers by an average of 1.2 points.

Lower variability compared to real responses

While capturing average consumer responses is important, it’s equally critical for synthetic data to reflect the diversity of opinions across individuals. Comparing distributions of both brand selections

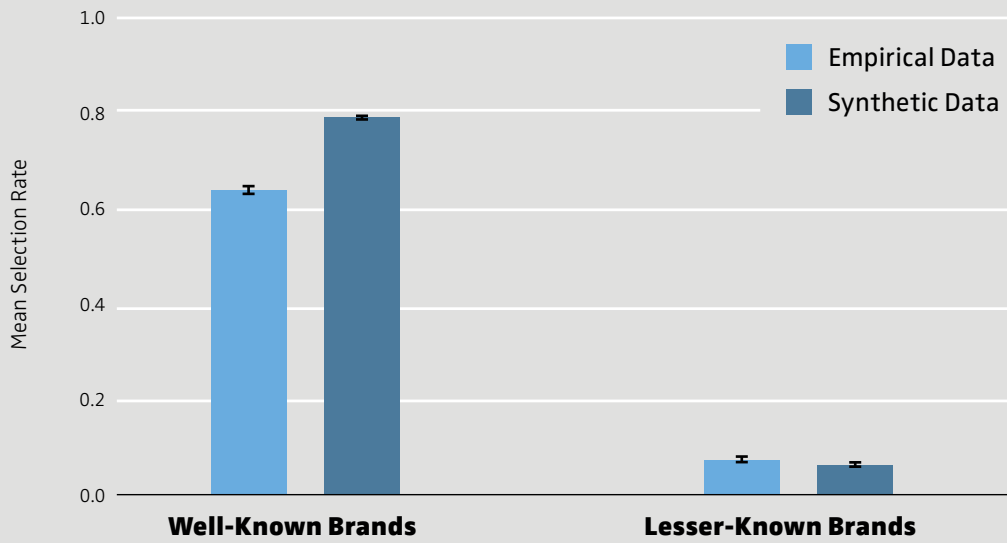
and numerical brand attitude ratings generated through direct Likert-scale elicitation, we found that synthetic responses were significantly less variable than real responses, particularly for well-known brands. Overall, this suggests that LLM-generated data tends to produce a more uniform pattern of responses compared with the natural diversity observed among human participants.

Semantic similarity ratings improve realism

Comparing the distributions of brand attitude ratings from real respondents, direct Likert-scale elicitation and semantic similarity rating for our survey shows a clear pattern: Semantic similarity rating produces greater variability than direct Likert scale elicitation, indicating improved realism (see Figure 3). However, it still overestimates brand attitudes and exhibits less variation than real respondents. Both synthetic methods underestimate the variability observed in human ▶

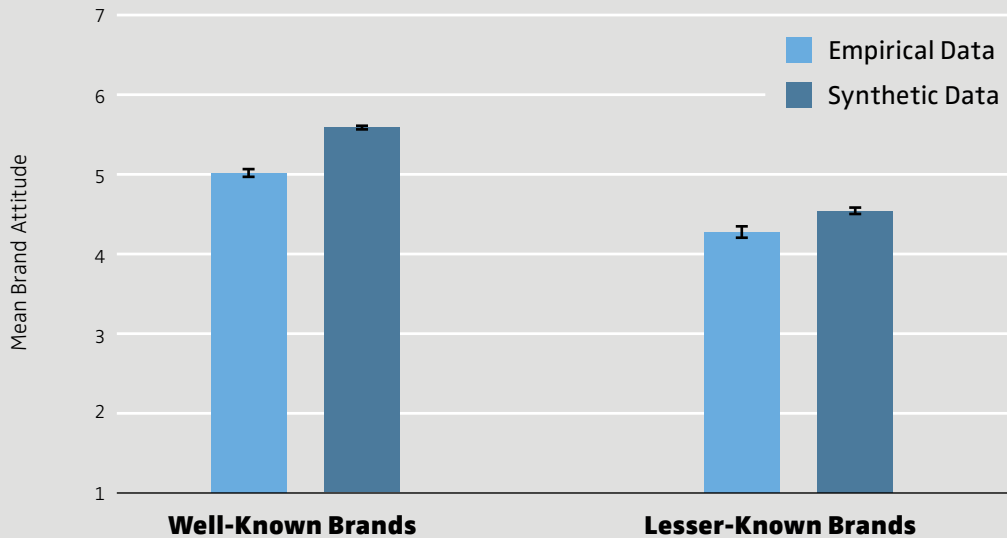
Comparison of brand selection results across types of brands between data from real and synthetic respondents

FIGURE 1



Comparison of brand attitudes across types of brands between data from real and synthetic respondents

FIGURE 2



responses, though semantic similarity rating shows partial improvement compared to direct Likert scale elicitation. In short, semantic similarity rating is an improvement, but not a replacement for real data.

Implications for market research: where synthetic data fits today

This study examined how well LLMs can simulate human survey responses using personalized “digital twins,” highlighting both their potential and current limitations for market research. LLMs tend

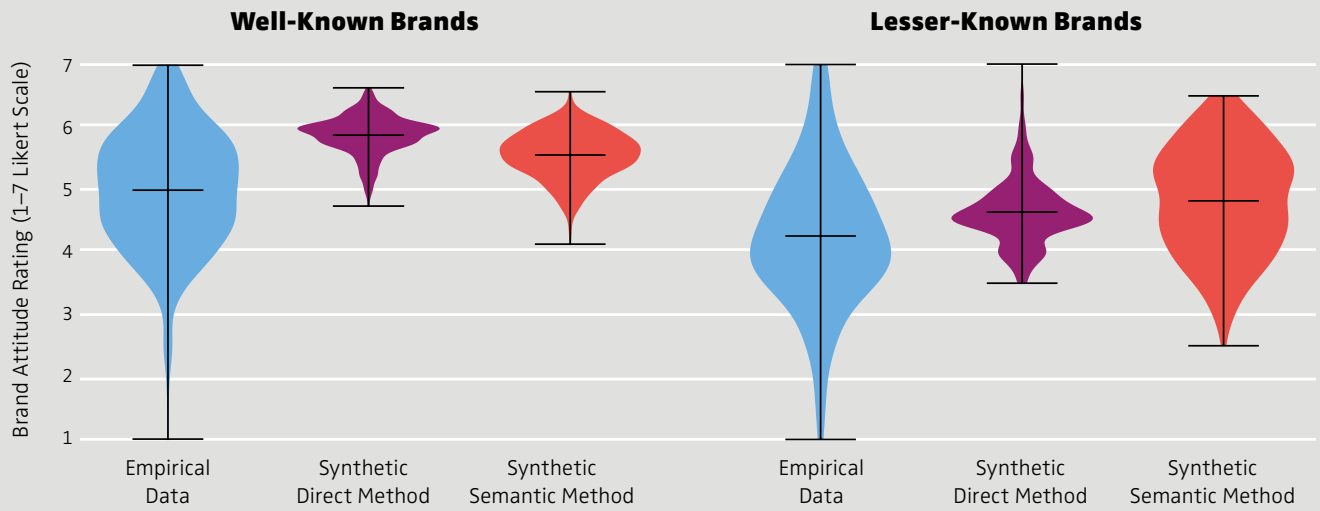
to overgeneralize or default to socially desirable responses, limiting their ability to reflect the nuance and diversity of real consumer opinions. As a result, synthetic survey data may currently be most suitable for early-stage concept testing or lower-stakes applications, rather than decisions requiring precise insights.

Looking ahead: improving synthetic survey accuracy

However, LLMs are advancing quickly, and new methods are emerging to address key limitations

Brand attitude distributions, comparing real responses with synthetic data generated via direct Likert-scale elicitation and semantic similarity rating

FIGURE 3



“
 Synthetic survey data may currently be most suitable for early-stage concept testing or lower-stakes applications, rather than decisions requiring precise insights.
 ”

in synthetic survey data. One promising direction is better answer-elicitation techniques, such as using the semantic similarity rating we present in Box 1, where the model generates a natural language response that is later mapped to a Likert scale. This plays to the model’s strength in text generation and can produce more human-like judgments. Further improvements may come from incorporating richer individual-level information, such as psychological traits or behavioral patterns, and integrating external context like product reviews or economic indicators, using retrieval-augmented generation (RAG). Together, these advancements could make synthetic data more realistic and representative of real consumer attitudes.

Balancing efficiency with ethics

For marketers, AI offers exciting possibilities: It can streamline research processes, reduce costs and quickly generate large datasets. Yet caution is needed, as biases and overgeneralizations can skew results and potentially mislead strategy. For consumers, AI could enable more agile, responsive marketing, but oversimplification of preferences risks producing less personalized or even irrelevant offerings. For society, widespread reliance on AI in market research raises ethical concerns, including the reinforcement of stereotypes, marginalization of minority viewpoints and potential disruption of traditional research roles. Careful oversight and new skill sets will be needed to ensure AI contributes responsibly to market insights without amplifying bias or inequality.

FURTHER READING

Kaiser, C., Kaiser, J., Manewitsch, V., Rau, L., & Schallner, R. (2025). Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 82–86). Association for Computing Machinery. <https://doi.org/10.1145/3708319.3733685>

Maier, B. F., Aslak, U., Fiaschi, L., Rismal, N., Fletcher, K., Luhmann, C. C., Dow, R., Pappas, K., & Wiecki, T. V. (2025). LLMs reproduce human purchase intent via semantic similarity elicitation of Likert ratings. *arXiv:2510.08338*. <https://arxiv.org/abs/2510.08338>