



THE EVOLUTIONARY ADVANCE OF SYNTHETIC RESPONDENTS

Ensuring Viability, Testing for Fitness

How *Mimetic* Are ‘Synthetics’?

The promise of synthetic respondents holds enormous appeal for the insights industry. If well-executed, it implies an infinite supply of tireless survey task rabbits, able to stand in for reluctant or scarce, sometimes unreliable, real-life consumers. Numerous synthetic data firms are waving this banner, making use of different methodologies to generate respondents.

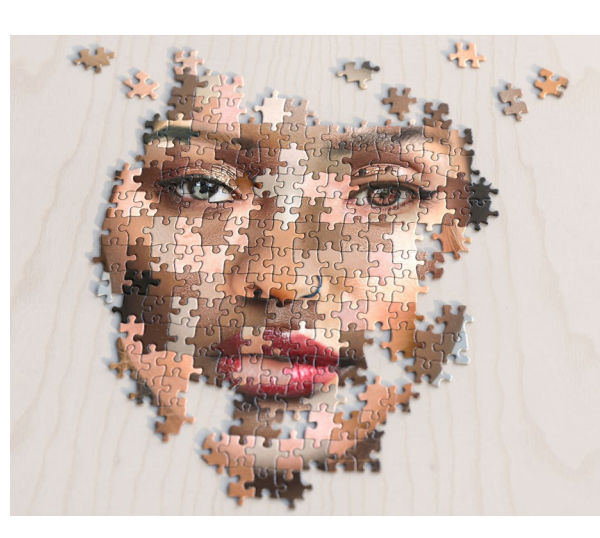
The potential for cost and time savings is impressive, but there are serious risks associated with trusting synthetic respondents until we have a clear grasp of the data quality implications and the tradeoffs we might be making. Rigorous vetting will be critical to identify usable sources of synthetic respondents, understand how well they perform at their best, and fully assess their limitations. Unfortunately, *most of the empirical assessments made thus far are missing the mark. This article explains why and what’s needed to fix the problem.*

First, What Are We Even Talking About?

Before going further, it’s important to clarify what we mean by “synthetic respondents”, since a stampede of vendors and other stakeholders have kicked up dust around the vocabulary in this field. Depending on the conversation, synthetic respondents might be referencing “gen pop” panels derived from post-trained LLMs or a more advanced product: specialty samples of so-called digital twins, groomed by feeding the models a large amount of customer and sector-specific data. While the primary focus here is on synthetic panels rather than digital twins, it’s fair to say that the principles below apply to both.

And What Are We Actually *Testing*?

Thus far, a key challenge for those trying to kick the tires has been the limited practical access to purpose-built commercial sources of synthetic respondents. In most of the independent evaluations we’ve seen, *the test synthetics were generated by the researchers themselves using some general-purpose LLM.* That’s because vendors are not generously sampling their wares for independent assessment and samples of comparable quality are too complex and costly for researchers to produce themselves. But expediency has consequences, no matter how well justified. *Many of the tests being performed on homemade, ready-to-hand, synthetic respondents are almost preordained to produce evidence of inadequacy.*



This is not to say that the most advanced synthetic respondent platforms are necessarily performing at a much higher level than published research would lead us to expect. Synthetic respondents are still in their evolutionary infancy, in need of more data and better training to produce survey avatars with greater human authenticity. But if the industry is to assess the state of the art accurately through that progress, we need the best tools for doing that. We don’t want to underestimate a promising technology simply because it flunked the wrong test.

The Origins of Synthetic Lifeforms: Why Perfectly Good LLMs May Grow Up to Be Bad Survey Respondents

To understand the root of the problems associated with the typical approach to testing synthetic respondents, we need to take a step back and consider how LLMs are trained, why many are not natively equipped to produce synthetic panels, and thus why testing the respondents produced by general-purpose LLMs won’t help us assess the kinds of specialty products that will ultimately be available to the industry.

Most approaches to synthetic respondent creation start by prompting an LLM with demographic/psychographic information (“You are a 45-year-old suburban mom with \$100k in household income...”) and asking it to answer questions as that persona. The feasibility of this idea rests on the premise that LLMs, being trained on vast amounts of human-generated text, have “absorbed” and internalized how different types of people think and can therefore sample from the vast distribution of verbal possibilities when prompted. How well an LLM can role-play human respondents depends on what the model is actually doing when prompted in this way—meaning that to understand what we might expect of them and how to test them, we must understand how these models are built.

Most general-purpose LLMs such as the Claude, Gemini and GPT series of models, are trained in a way that ultimately makes them poor imitators of a human survey taker. They are taught to be obliging, line-toeing assistants, not independent thinkers, because that was the destiny their developers foresaw for them. The problem of “sycophancy” has been much discussed, even in the popular press. The root of that problem can be explained as follows:

The first stage in training an LLM produces a “base model,” which is a pure reflection of the distribution/predictor of the text it was trained on. At this point, the model can effectively simulate any kind of text, including plausible survey responses. After all, a “very likely” response, or even “4 on a 1-5 scale” is the sort of thing a model has become familiar with when trawling for correlations in all the human text its trainers can find to feed it. However, given the variety of text the model has learned to imitate, the practical value of a base model is still limited: it will be prone to going off topic, outputting nonsense, or otherwise producing text that isn’t what the user is looking for.

To address these issues, model developers then engage in various post-training steps: fine-tuning for conversational coherence, reinforcement learning with human feedback (RLHF) to teach the model to create “helpful, honest and harmless” outputs, and more recently reinforcement learning on reasoning, to solve difficult problems or generate reliably correct code. While these steps have driven the transformation of AI from an interesting toy to a now nearly indispensable tool used by millions daily, post-training has the side effect of driving the model into a particular persona, referred to within the industry as the “assistant” persona. We’d logically expect this persona to provide middle-of-the-road, slightly positively skewed answers to survey questions. When someone in the industry tests synthetic respondents using models like GPT-4o, that tends to be exactly what they find. This phenomenon can be observed broadly in model training and needs to be factored into every sort of assessment we make about a model’s fitness-to-purpose.

Model Evolution and True Fitness: We Still Can’t Test the Performance of Synthetic Panels Using Post-trained LLMs

Over the past two years, progress in model development has been remarkable on all fronts, but synthetic respondent vendors and those who evaluate their wares often continue, for very practical reasons, to use older models that keep costs down. *This is in fact a reasonable choice: while newer LLM models are doing a much better job with the targets they aim to hit based on the current post-training process, those models are still not aiming for the targets that matter to companies creating synthetic respondents.* As a result, we wouldn’t expect them to better impersonate the average human off the street. And, in fact, tests that [used newer models like GPT-5.2](#) have failed to find improvement vs older generations, just as we might have predicted. That means they cannot generate truly “fit” test synthetics for purposes of vetting purpose-built panels.



At the same time, commercial panel companies are further fine-tuning base models by infusing large quantities of survey data to elevate the product, turning them into “expert survey-takers”. It’s still early days, of course, but initial releases look promising, and by implication, this progress will further compromise evaluations by producing greater separation between commercial panel synthetics and the home-made variety grown from general-purpose LLMs that independent researchers tend to rely on. Meanwhile, commercial panel vendors are not publishing the kind of exhaustively rigorous test results one would need in order to fully understand performance. That burden has been left to clients and the industry to accomplish, largely at their expense.

If we want to lay sharp bets on the role for synthetic respondents, we need to do it right by testing the highest caliber of synthetic respondents we can find, using methodologies that produce valid measurement of their performance. It may be pie-eyed optimism, but we are hoping that major synthetic panel vendors will make their respondents more available for cost-effective study so that researchers can use the right models for the job. We urge vendors who are selling synthetic data to embrace and facilitate efforts that will, long term, make synthetic data a regular and trusted source of insights in an industry running short of human respondents.

Postscript: What’s Next at NAXION

NAXION’s ongoing research-on-research program is testing how close the synthetics from commercial panels come to humans both in broad terms and in nuanced particulars. Equally important, our research aims to pinpoint what the specific shortcomings and limitations we observe may tell us about the statistical techniques we can legitimately apply to wring full value from the data. We will be sharing the results of our efforts over the coming months.

About The Author



Hunter Geisel
Group Director, NAXION
215.496.6862
hgeisel@naxionthinking.com

Hunter’s expansive role at NAXION includes responsibility for developing and testing AI applications throughout the data development workstream as well as research design and consulting to clients on business-critical questions. He collaborates closely with internal stakeholders and with NAXION’s clients to harness the power of AI responsibly, leveraging creative talent and technical expertise to engineer standard processes and devise custom solutions. He has a Bachelor of Science degree in economics from the University of Maryland.

About NAXION

NAXION is a nimble, broadly resourced boutique that relies on advanced research methods, data integration, and sector-focused experience to guide strategic business decisions that shape the destiny of brands. Our century-long history of innovation has helped to propel the insights discipline and continues to inspire contributions to the development and effective application of AI to research and data science techniques. For information on what’s new at NAXION and how we might help you with your marketing challenges, please visit <https://www.naxionthinking.com/>.