NAXION
thinking

THE EVOLUTIONARY ADVANCE OF SYNTHETIC RESPONDENTS

Ensuring Viability, Testing for Fitness

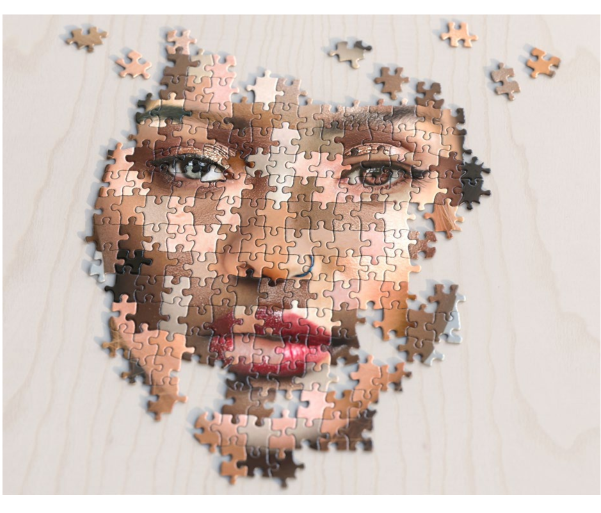
How *Mimetic* Are ‘Synthetics’?

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. With the explosion of interest in AI over the past several years, an array of synthetic data vendors is rapidly emerging, 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. Gaining this understanding is made challenging by the fact that current methods of synthetic sample generation vary both in their general ability to mimic human survey-takers and the specific claims they can legitimately make. Rigorous vetting will be critical to identify usable sources of synthetic respondents, understand how well they can perform at their best, and fully assess their limitations.

Many stakeholders from different corners of the insights industry – commercial customers and research practitioners, traditional panel vendors, and academics – have all joined the conversation, some with theoretical critiques, some with empirical research on synthetic data outcomes. *Most of the empirical assessments may be missing the mark. This article explains why.*

First, What Are We Even Talking About?

Before going further, it’s important to clarify what we mean by “synthetic respondents”—a term clouded by ambiguity at a moment when blazing innovation and a stampede of stakeholders are kicking up dust around vocabulary. Depending on the context or conversation, synthetic respondents might refer to digital twins created from a limited number of qualitative respondents; digital twins created from large customer databases and fed lots of personal data; or gen pop panels designed from LLMs—with or without supplemental survey data to enrich them. *Our focus here is on the last – synthetic gen pop panels rather than digital twins – although many of the issues to consider have bearing on all of them.*



And What Are We Actually *Testing?*

Thus far, a key challenge for those trying to kick the tires in a serious way has been the limited access to commercial sources of synthetic respondents for rigorous vetting. In most evaluations we’ve seen, *the synthetic respondents tested were generated by the researchers doing the evaluation* because, for the most part, vendors are not generously sampling their wares for purposes of systematic assessment. Since the most promising methods of generating high-quality synthetic panels are highly complex and expensive to prepare, there can be a critical mismatch in methodologies. As a result, many of the tests being performed with ready-to-hand or “homemade” synthetic respondents are almost preordained to produce evidence of failure.

That is not to say that these synthetic respondent platforms are necessarily performing at a higher level than suboptimal testing is able to discern. We already know that many implementations of synthetic respondents are incapable at the moment of handling all the question formats we might use with human respondents (e.g., constant sums and numeric responses) and there are clearly other limitations, including synthetic respondent memory and/or length of attention. And we also know that synthetic respondents are still in their evolutionary infancy, in need of more data and better training to produce truly dimensional avatars with greater human authenticity.

It’s fair to say, then, that the problem straddles both sides of the aisle: the people who build mimetics are not always using the best models for the purpose and the people testing mimetics aren’t either. Risk of failure is built in at both ends. In order to avoid dismissing a promising technology simply because it flunked tests that were improperly constructed at an early stage of evolution, we must first take another step back to consider how models are trained and the implications both for downstream development of an application like synthetic panels and for the way we test them.

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

Most approaches to synthetic respondent creation start by prompting a general LLM model with demographic/psychographic info ("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.

It turns out that most commercial 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-towing assistants, not independent thinkers, because that was the destiny their developers foresaw for them. It turns out that towing the line can mean hugging the mean.

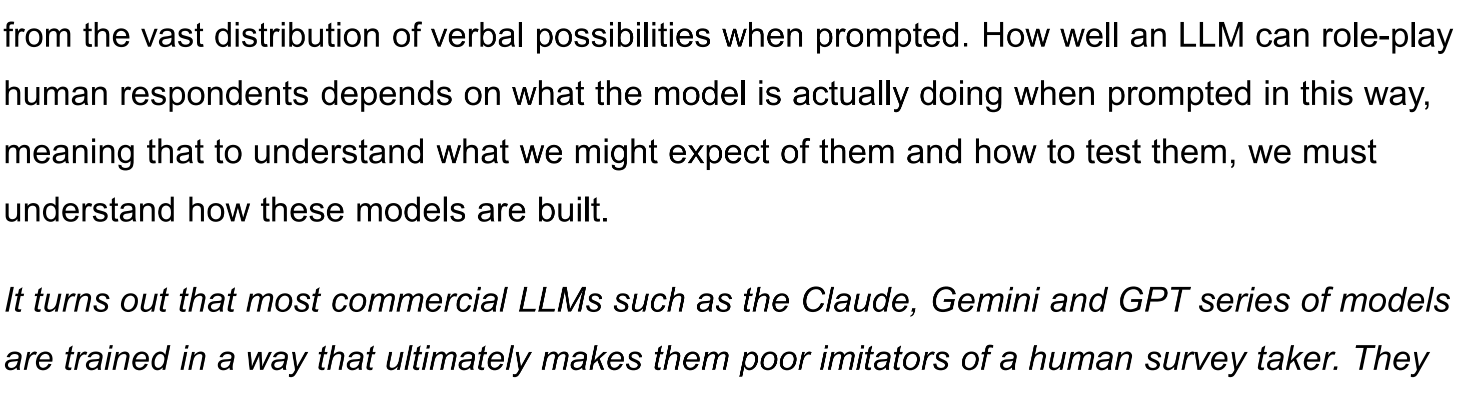
The first stage in training an LLM produces a “base model,” a pure reflection of the distribution/predictor of the text it was trained on. At this point, the model can effectively simulate any kind of text, potentially including survey responses. After all, a “very likely” response, or even “4 on a 1-5 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 in the types of text the model has learned to imitate, the practical value of a base model is limited: it will be prone to going off topic, outputting nonsense, or otherwise producing text that isn’t what the user is looking for.

For reasons of practicality and safety, 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. These steps have driven the transformation of AI from an interesting toy to a tool used by millions of people daily. However, they have the side effect of driving the model into a particular persona, referred to within the industry as the “assistant” persona. Anthropic recently released a report that includes an examination of this persona, finding that it skews helpful, methodical, and calm, in the style of a consultant or teacher. When taking a survey, we’d naturally expect this persona to provide middle-of-the-road, slightly positively skewed answers. 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. Training the model to act as a certain sort of character generalizes to all of its outputs and introduces biases that may not be immediately obvious. As one surprising example, OpenAI explored the opposite of its usual training process in its paper on emergent misalignment, where fine-tuning the model to make it behave badly in one way caused it to generate “harmful” outputs in totally unrelated contexts.

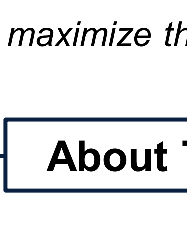
Model Evolution and True Fitness

Over the past two years, progress in model development has been remarkable on all fronts, but, understandably, synthetic respondent vendors and those who evaluate their wares often continue to use older models to keep costs down. *The problem is that while newer 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 the newer, smarter models to better impersonate the average human off the street**. And, in fact, tests that have used new models like GPT-5.2 have failed to find improvement vs older generations, just as we might have predicted.



While generating truly mimetic respondents from post-trained models has little hope of success, improvement is still within our grasp. A more sophisticated approach – using a base model as a “general simulator” that is then fine-tuned on large amounts of diverse survey data – may do a better job of replicating the distribution of human survey responses. The same penchant for generalization that makes assistants dysfunctional would instead work in favor of such a model. The key hurdle is the size of the commitment required, including very significant computing resources, model expertise, and vast amounts of data. Ultimately, though, there’s no help for that. *If we want to lay bets on the role for synthetic respondents while ensuring we don’t lose everything in the gamble, we need to do it right: generate synthetic respondents using techniques that maximize the chance of success and apply valid test methodologies that maximize the chance of measuring it.*

About The Author



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

Hunter is a Group Director at NAXION whose expansive role includes research design and consulting to clients as well as responsibility for developing and testing AI applications throughout the data development workstream. He collaborates closely with internal stakeholders and with NAXION’s clients to harness the power of AI responsibly, leveraging creative talent and and technical expertise to create custom solutions for business-critical objectives as well as engineer standard processes. 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/.

Winter 2026

NAXION
thinking