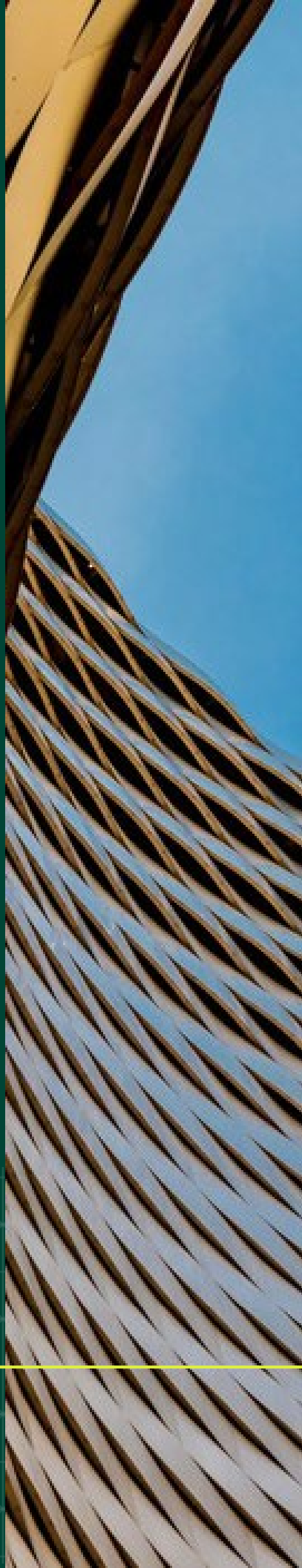
Part 1

Don't replace your human sample just yet:

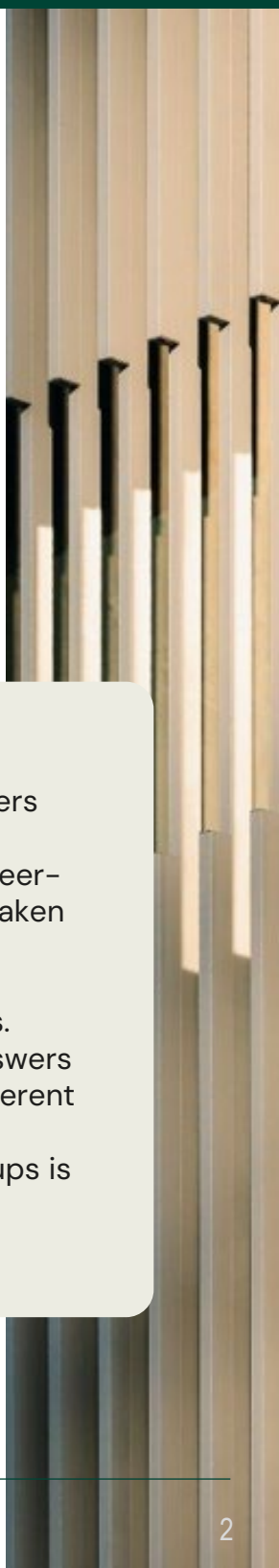
An evidence-based guide to synthetic respondents



THE DIRECTIONS group

The question has officially gone mainstream:

Should we be using synthetic respondents in research?



Whether you lead an insights team, drive marketing strategy, or steer innovation, you are likely feeling the pressure to conduct research faster and cheaper.

Synthetic respondents (aka, AI models programmed to answer survey questions as if they were target customers), are being offered up as the solution.

It is a tempting pitch. Some vendors claim they can instantly fill hard-to-reach quotas and replace 15 to 85 percent of the real people taking your surveys with "synthetic respondents." Eventually, they say, this will fully replace real people. And the promise is to cut fieldwork costs by 40 percent. However, this framing is aggressive and ahead of what the evidence supports.

At the same time, serious researchers are pushing back. They are raising foundational questions, based on peer-reviewed critiques, and should be taken seriously. They argue there is no theoretical reason for naive LLM respondents to be valid as samples. Because human answers and AI answers are generated in fundamentally different ways, running standard statistical comparisons between the two groups is technically meaningless.

So, who is right about synthetic respondents?

Both sides cannot be entirely correct, and a research partner who tells you otherwise is not serving your interests. Moving ahead with synthetic respondents without thorough consideration can risk injecting hidden bias into your data. You could end up basing major strategic decisions on fiction.

The search for an answer

Instead of picking a side or staking claim to the middle ground, we sought a thoughtful answer. We spent the past year running pilots, reviewing academic research, and putting synthetic respondents head-to-head with humans in a live study to see exactly where they help, and where they fail.

In this article,

We answer questions leaders are asking about synthetic respondents:

- Can synthetic respondents actually replace human survey takers?
- Is synthetic data a new concept in market research?
- What are the risks of using AI models as survey respondents?
- How do Large Language Models (LLMs) create contamination and bias in survey data?
- Should AI tools replace or amplify human intelligence in market research?

Data modeling isn't new, but its history can inform how to use it in the future.

One of the most important things to understand about synthetic data is its age. The underlying idea traces back to the 1940s, when scientists Ulam, von Neumann, and Metropolis used early simulations to solve complex physics problems.

In 1987, a researcher named Donald Rubin created a formal way to fill in missing survey data with plausible modeled answers. By 2011, using complex math to fill in data gaps became a standard tool in epidemiology and official statistics. In 2014, deep generative models opened an entirely new era of simulated data.

AI-powered synthetic respondents, which hit the commercial market around 2022, are just the latest chapter in this story. Decades of academic discipline have already shown us how to use modeled data with rigor. That lineage matters because it tells us what is possible and where the boundaries lie.

What is genuinely new (and more complicated) about today's LLMs?

Large language models expand what synthetic data can do. They produce coherent, open-ended rationales instead of just numbers. A demographic backstory in a prompt can be deployed thousands of times in minutes. A single model can span consumer goods, healthcare, and financial services, instead of needing a custom mathematical model for every single study. But this expansion comes with new risks.

What are the risks of using LLMs in research?

Data contamination.
The model may have already been trained on the very category you are studying. It is built on past data, meaning it is often just recalling public sentiment rather than predicting how a consumer will react to a totally new launch.

The plausibility trap.
Fluent, well-written synthetic responses create a false sense of confidence. Just because an answer sounds true does not mean it is.

Flattening of nuance.
Models tend to regress to the "typical" response, hiding the outliers and flattening the nuance where true human insight usually lives. If you ask a model to rate something on a scale of 1 to 5, it clusters its answers around the middle. This is a risk because the opinions on either end are often where the most valuable human insights live. Instead of capturing how people actually think and feel, flattening leaves you with an artificial, middle of the road consensus.

Amplifying bias.
Generative models replicate the biases in their training data, meaning underrepresented groups often suffer the most.

Opaque sampling.
With older math models, researchers could describe what population they were sampling from. With an AI model, we cannot.

The Directions Group position on synthetic respondents

THE DIRECTIONS group

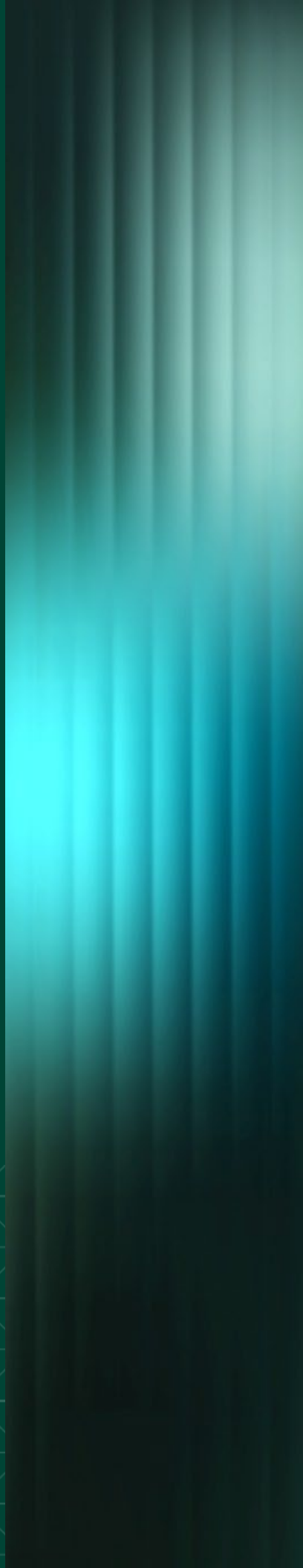
Instead of tossing out AI or going all in, we are asking better questions about when it is appropriate to use it. Any technology that powers our work must serve to answer our clients' business questions in the best way possible.

Our core position remains unchanged: Synthetic tools should amplify the best of human intelligence.

We believe synthetic respondents are a real tool with real limits. Anyone can access AI resources, but AI cannot replicate human empathy. Do not let algorithms replace what only human respondents can answer.

Part 2

What our data showed when we tested synthetic vs. human respondents in a brand survey



In this article

We share evidence about the accuracy of synthetic data, and answer the following questions:

- How accurate are synthetic vs human respondents?
- Can AI accurately measure brand health and favorability?
- What happens when AI takes a live brand tracker survey?
- Can you apply a math formula to fix AI data inflation?
- Does the way you ask an AI model a question change the result?

As the debate over AI in market research heats up, leaders are asking us a critical question: What is the evidence for using synthetic respondents in research?

To find a substantive answer, we put synthetic vs human respondents in the same survey.

In Q4 2025, The Directions Group ran a head-to-head comparison embedded in our QSR (Quick Service Restaurant) Brand Momentum tracker, our ongoing fast-food brand health survey.

This flagship study covers seven major brands. We ran the survey with 934 human respondents and 298 synthetic respondents.

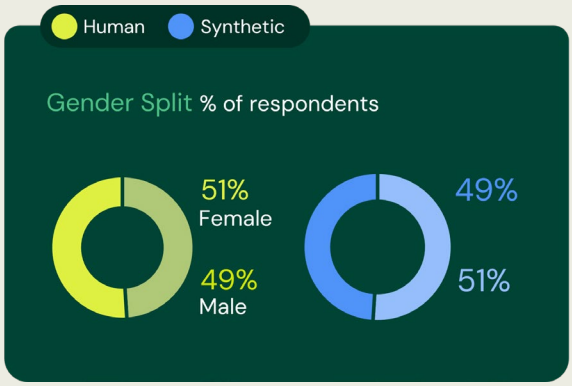
The study gave us a direct, side by side comparison across awareness, favorability, and recommendation measures.

Here are the results.

Finding 01. Demographics Match

On basic structural questions, synthetic and human respondents came in nearly identical. The female to male split was 49 to 51 for synthetic versus 51 to 49 for human respondents. Age distributions tracked within two to three percentage points across almost every age group.

This is the result vendors lead with, and the data supports it. Synthetic respondents can be demographic mirrors of your target population. Demographic similarity tells you the sample looks right, but it does not tell you the sample thinks right.



Finding 02. Rank Order Mostly Holds. Absolute Levels Do Not.

When we looked at a core metric like "Definitely Would Recommend," the brand tracker synthetic data unraveled.

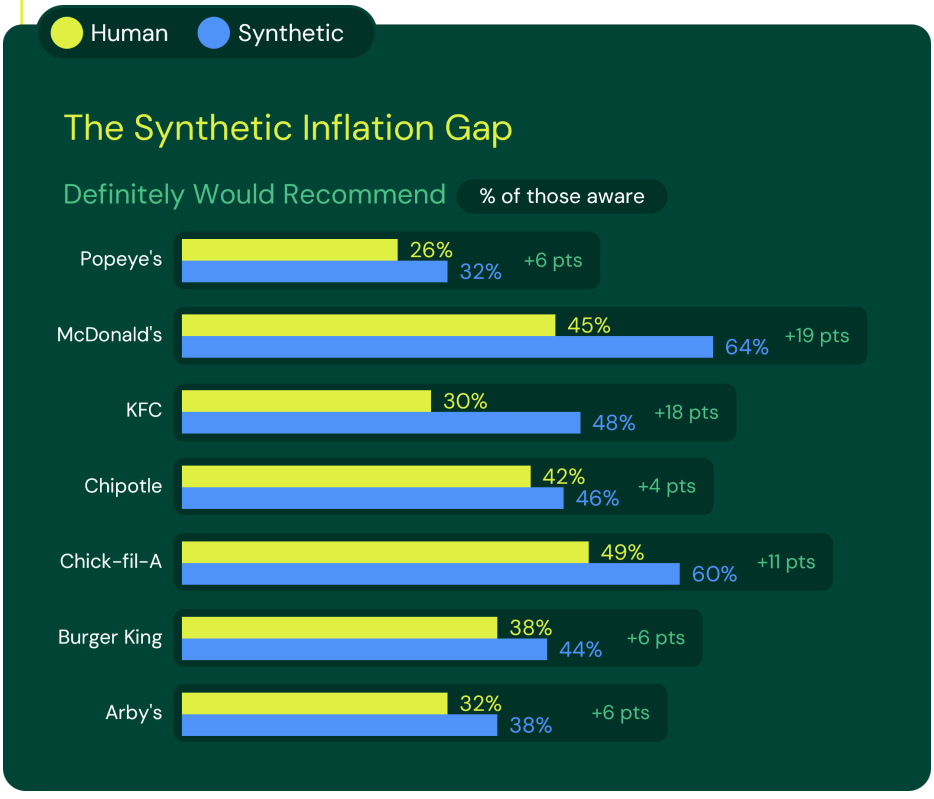
Synthetic respondents inflated top-box scores (the highest possible ratings) by 4 to 19 points across all seven brands studied.

The inflation was not uniform.

- Chipotle was inflated by 4 points.
- McDonald's and KFC were inflated by 19 points each.

The brands with the largest AI bias were also the brands with the most online discourse. This is what you would expect if the AI was distorting its answers with public sentiment rather than simulating a real consumer decision. Rank order was partially preserved.

Chick-fil-A leads in both samples. KFC and McDonald's switched places when measured synthetically. That is a highly meaningful error if competitive positioning is the purpose of your study.



Perhaps most importantly, there was no reliable correction factor. Because the inflation varied five times across brands, a researcher cannot just subtract a standard offset to recover the human number.

A claim such as "rank order is preserved," (offered by most synthetic-respondent vendors) needs to be qualified: it is preserved at the top but not stable across the middle of the distribution.

Finding 03. The method matters more than the model.

Our findings align with the most credible academic validation of synthetic data accuracy to date. In 2025, a study run by PyMC Labs and Colgate-Palmolive tested AI-generated consumers against 9,300 human responses across 57 personal-care concept tests (tests of new product ideas). That study found that synthetic respondents achieved an impressive 90 percent relative to human product rankings.

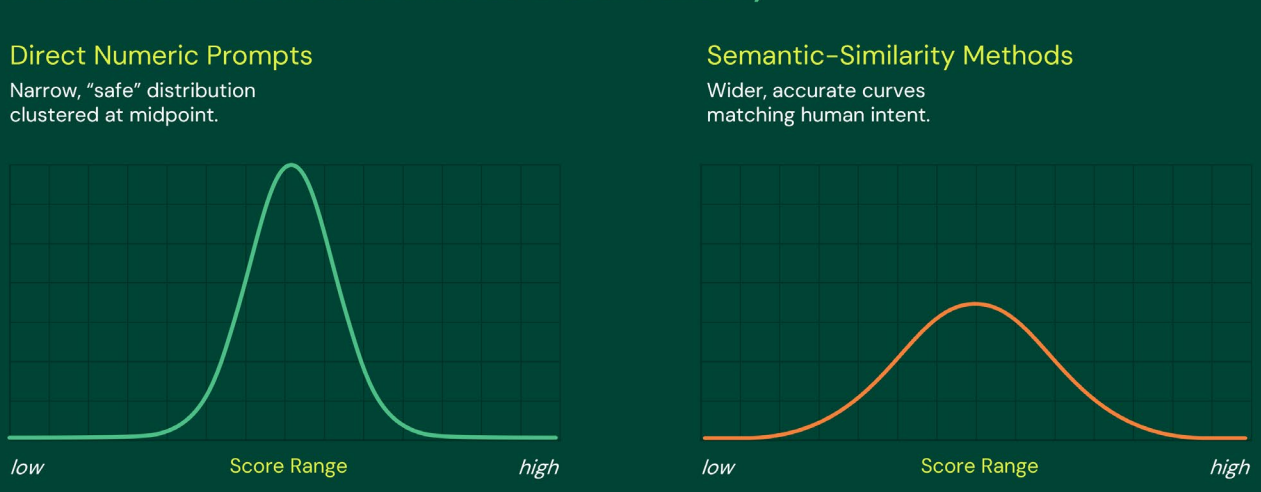
Those results required a highly specific method. When the researchers used direct numeric prompts (like asking the AI to rate something on a scale of 1 to 5), it produced safe, middle-of-the road answers, which replicated the findings other critics have identified.

LLMs reproduce human purchase intent when you ask them right.

To reach 90% relative accuracy, researchers had to ask the AI to write a free text response first, and use advanced natural language processing methods to map that text back to a rating scale (the semantic-similarity method). However, this method only works in domains where the LLM's training corpus is rich, when the question is rank-ordering preferences, and the questions themselves are done well.

In the research, the 90% level of accuracy was demonstrated in CPG personal care, and was not validated in B2B, healthcare-HCP, or low-incidence audiences.

Elicitation Method Dictates Data Validity



What this means for your research.

If you are considering synthetic respondents for studies where relative order is the primary output, the evidence is genuinely encouraging.

For tasks such as narrowing 8 concepts down to 3 before launching full-scale fielding (a complete survey), or ranking 30 product features, there are real use cases with real support.

If you need absolute levels, synthetic data is simply not a substitute for real people.

If you are reporting brand awareness percentages, favorability scores, or recommendation rates as business metrics, our data makes the case clearly. The inflation is real, it is systematic, and it cannot be easily corrected.

The research profession deserves this level of honesty about a technology that vendors are actively selling as a time and cost saving solution. So, as we continue testing, we will continue publishing what we find.

To discuss how these findings apply to your specific research design, feel free to contact the Directions Group Centers of Excellence.

We help clients evaluate vendors, scope pilot designs, and put validation in place before any insight is finalized.

Part 3

Synthetic Data Use Cases:

Where AI survey takers are helpful and where they are harmful

In this article

We answer critical questions leaders are asking about AI survey takers.

- How do you draw a safe line between where synthetic respondents help and where they are harmful?
- Why is AI accurate at putting product options in the correct order, but unreliable at predicting stand-alone scores?
- What are the six specific research tasks where real-world evidence supports using simulated survey respondents?
- What are the four areas where using AI-generated data puts your business decisions at risk?
- What seven questions should you ask any research vendor before agreeing to use synthetic data?

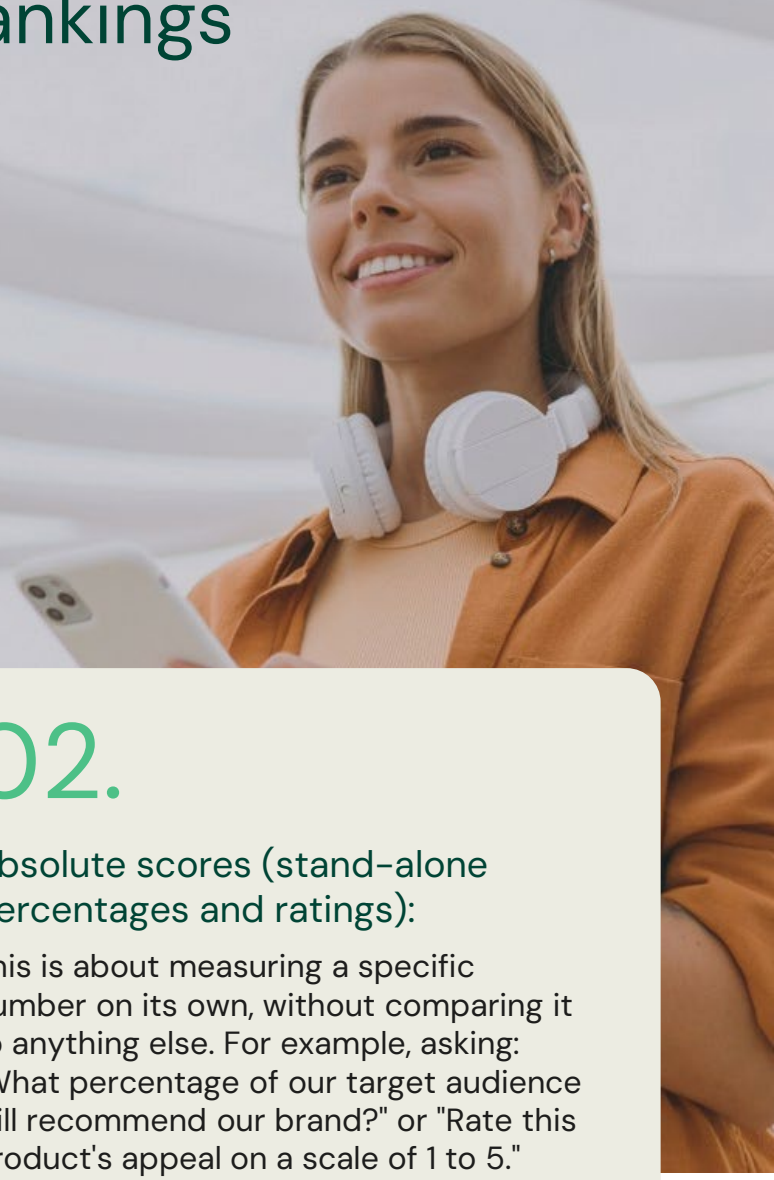
Business leaders need a practical way to evaluate synthetic respondents before putting their research (and their decisions) at risk.

We have spent the last year reviewing evidence and running our own studies so that real-world validation dictates where AI tools belong in research design. This level of discipline is essential for achieving decision readiness, because a study only succeeds if it provides business teams with enough reliable evidence to launch a product, change brand positioning, or invest capital with confidence.

The Dividing Line

Comparing relative rankings and absolute scores.

The single most useful distinction we have found is this: AI-generated respondents perform best when relative order matters, but they fail when you rely on absolute, stand-alone scores.



To understand this distinction, it helps to look at how these two types of data work in practice.

01.

Relative rankings (putting options in order of preference):

For example, asking: "Does a consumer prefer Product A over Product B?" or "Rank these five new flavor ideas from favorite to least favorite." The evidence shows AI is good at putting options in a meaningful sequence. A landmark 2025 study by Colgate-Palmolive and PyMC Labs found that AI-generated respondents achieved a 90% match to real human concept rankings. Similarly, our own Q4 2025 tracker found that the general ranking of fast-food brands was mostly preserved.

02.

Absolute scores (stand-alone percentages and ratings):

This is about measuring a specific number on its own, without comparing it to anything else. For example, asking: "What percentage of our target audience will recommend our brand?" or "Rate this product's appeal on a scale of 1 to 5." When we look at these stand-alone metrics, AI-generated respondents consistently show inconsistent value—sometimes inflating, sometimes deflating, and in most cases flattening out the value. For example, in our Q4 2025 tracker, the AI inflated the top recommendation ratings by 4 to 19 points across seven major restaurants.

In short, ordering and screening tasks tend to work. Trying to rely on stand-alone, absolute scores to make a business projection does not.

Five strategic use cases for synthetic respondents, supported by evidence.

There are five specific areas where the empirical evidence supports using synthetic respondents.

01. Sorting large lists of features and claims

Ranking features, claims, or messages where only the hierarchy matters.

For example, if you need to quickly understand which of 30 concepts deserves a real study, simulated respondents can shorten a prioritization exercise from weeks to hours.

02. Concept Pre-Testing

Screening 8 concepts down to 3 before committing to full-scale choice modeling or monadic fielding

Pre-testing helps teams avoid chasing a bad idea. Simulated survey takers let you weed out the duds early for a fraction of the cost. The evidence suggested this could be done, along with validation against real people.

03. Survey Instrument Pre-Tests

Sanity-checking question wording, flow, and skip logic before your study launches.

Simulated survey takers can quickly point out confusing questions, items that accidentally ask two things at once, or broken paths in your survey. As a result, you can catch glitches and clean up the questionnaire before a real panel ever sees the survey.

04. Directional Discovery

Initial exploration before a robust primary study

If you are asking "what questions should we be asking?", synthetic respondents can generate early hypotheses that sharpen the design of the human study that follows.

05. Hypothesis Generation

Generating qualitative rationales for why a concept might land with a target audience.

This is a qualitative benefit unique to large language models. It gives you the ability to produce coherent narrative responses that can inform creative and strategy work, with the understanding that these are hypotheses rather than final findings.

A hypothesis yet to be validated: Quota Boosting

Filling 15 to 20 percent of a hard-to-find subgroup

If you need to survey a very hard-to-reach group, such as specialized cardiologists or a rare type of consumer, finding enough real people can be slow and expensive. Suppliers are suggesting that synthetic respondents are a key use case (using AI to fill in gaps within the population), but historical evidence suggest that statistically speaking, this does not hold.

Four areas where simulated data does not belong.

Just as the evidence shows us where AI tools can assist, it is equally clear about where synthetic data should not be used to guide business decisions.

01.

Tracking your brand's performance over time

Our brand study showed that AI inflated favorability scores by 6 to 12 points, and recommendation scores by 4 to 19 points. This is not a minor statistical wobble. It is a predictable bias because the AI's "opinions" are heavily distorted by public sentiment and chatter it has already scanned on the Internet.

02.

Creating financial models and market size projections

Estimating sales volumes or product share requires highly accurate, real-world numbers. Simulated data consistently misses these targets, meaning it cannot be used to justify major capital investments.

03.

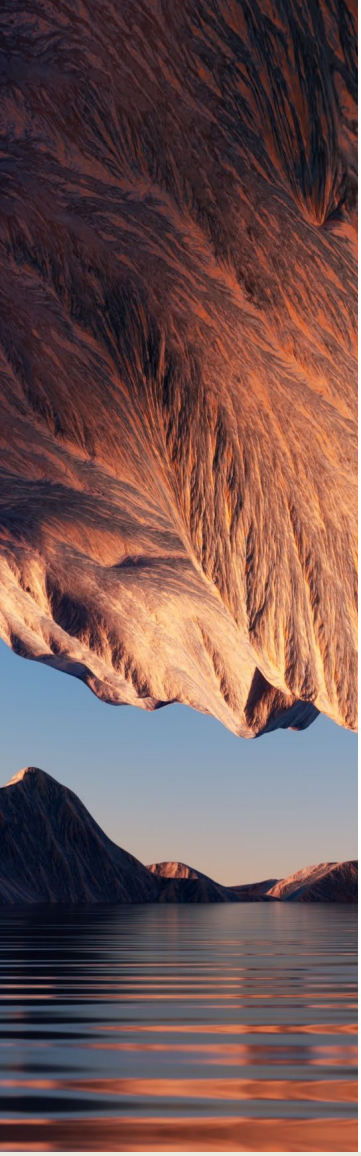
Making high-stakes, either-or decisions

We cannot assume the AI will get the basic direction of consumer reaction right. In independent tests of political polling, the AI got the direction wrong for about one out of every five people, predicting that an opponent of a candidate would actually support them.

04.

Sizing up rare, business-to-business (B2B) audiences

For an AI to successfully simulate a consumer, its underlying model must have already processed a massive amount of high-quality data about that specific group. For highly specialized professional roles or narrow business audiences, that training data is incredibly thin, making the AI's guesses highly unreliable.



The critique you need to take seriously.

Synthetic respondents have no theoretical reason to be considered true statistical samples of any meaningful population.

There is no built-in mathematical guarantee that an AI model's simulated answers will match what your specific customers think. Just because a vendor proves their model worked well on one study does not mean it will work on your next project.

The disciplined response to this challenge is not to ban AI from your research. It is to demand proof.

Every hybrid research design should include a small group of real human respondents to serve as a benchmark. You should never base a business decision on AI responses simply because they sound realistic and fluent. Fluency is not the same as accuracy.

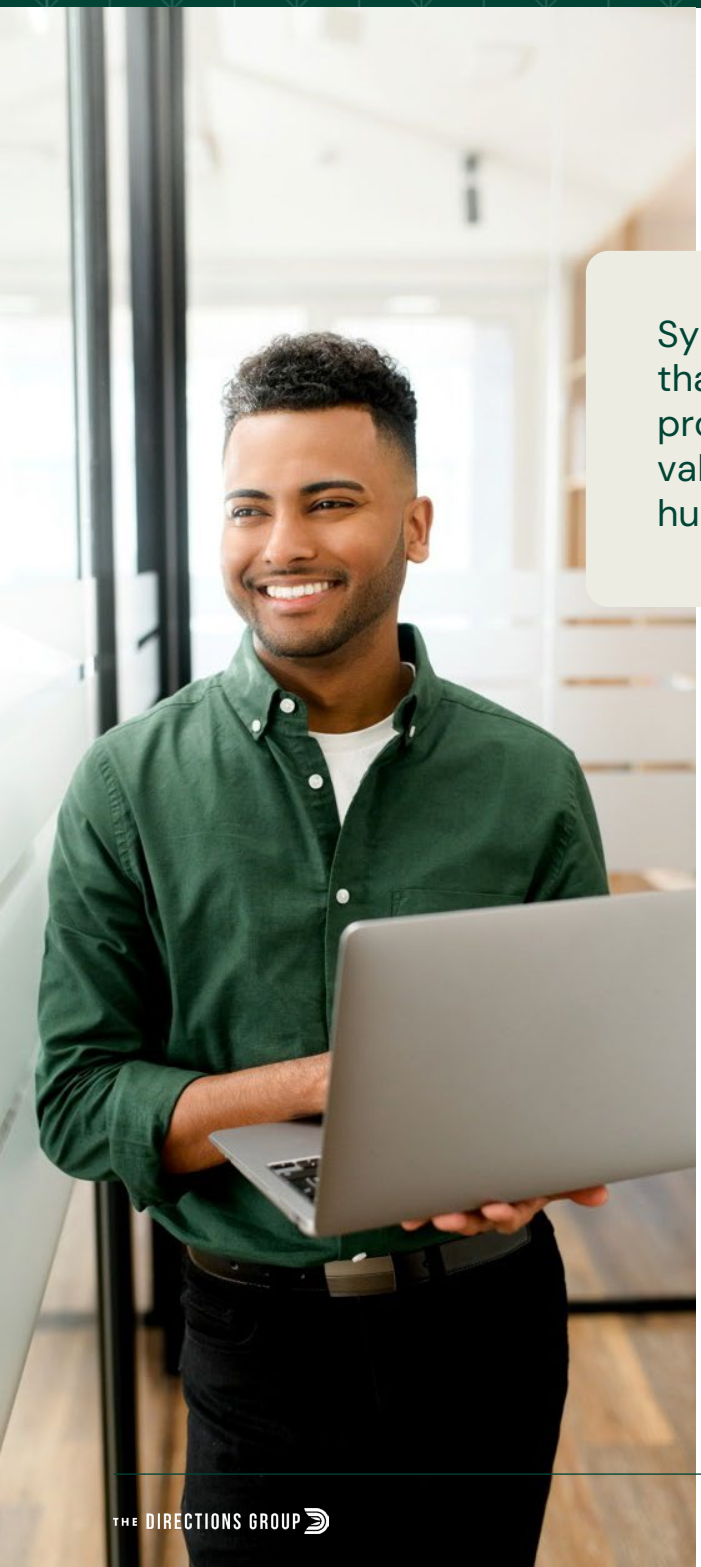
Seven Critical Questions to Ask Before You Deploy AI Survey Takers

If a vendor pitches you a study using synthetic respondents, we recommend asking these seven questions before you agree to the methodology. The answers will tell you if the data is safe for decision making.

- Will you set and write down clear accuracy targets for our specific study before we start?
- Has your AI model already been trained on the category or data we are trying to predict? (If it has, the AI might just be repeating old public sentiment rather than predicting how consumers will react to your new launch.)
- Where does your model perform well, where does it fail, and how do you prove it?
- When you update or change the underlying AI model, do results change?
- What tests has your model failed that you have not published?
- How is the persona prompt constructed, and on which demographics?
- What elicitation method is used? (Direct, follow-up, or semantic-similarity?) In other words, how do you get ratings from the AI? Do you ask for direct numbers or do you have the AI write text first?

The answers to these questions will tell you more about the credibility of a study design than any hand-selected success story.

The practical starting point.



The most honest summary we can offer?

Synthetic respondents are a legitimate tool that belongs in a well-designed research program—in specific roles, with documented validation, and never as a replacement for human respondents when the answer matters.

Scoping the role of AI correctly for your specific study is exactly the kind of work the Directions Group Centers of Excellence was built to do.

We help clients evaluate when synthetic data fits the business question, design the right hybrid research program, and put validation in place before any insights reach a deck.

THE DIRECTIONS group

If you'd like to discuss how these findings can apply to your research, feel free to contact the Directions Group Centers of Excellence.

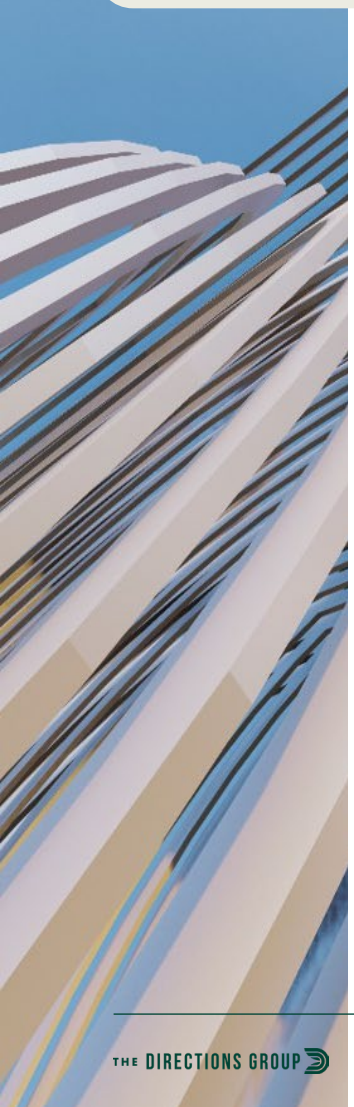
Part 4

Five principles for using synthetic respondents responsibly

In this article

We answer important questions to guide research design:

- What are the standards for using AI survey takers responsibly?
- Why must every synthetic deployment require a human benchmark?
- Why does fluent, natural-sounding AI text create an illusion of validity?
- Why is synthetic data accurate for relative ordering but unreliable for stand-alone absolute scores?
- Why does how you ask questions matter far more than which model you use?
- Why should you be cautious about flashy new methods?



Every new research method eventually requires a clear set of standards.

As synthetic respondents enter the market research landscape, the insights profession has reached that critical moment.

Over the past year, The Directions Group has studied the peer-reviewed science, run controlled head-to-head comparisons in live studies, and built an evidence-based approach to using synthetic respondents.

In the previous three articles, we looked at the eighty-year history of modeled data, the findings from our own restaurant brand tracker, and a practical decision framework grounded in evidence.

This final article in our series outlines the five core principles we apply to any study using synthetic respondents. These principles establish a standard of rigor that protects your research metrics and ensures decision readiness.

Principle 01.

Amplify, don't replace.

Synthetic respondents serve human researchers. They should support our work, not speak for the real consumers we are studying.

Surveys are an act of motivated communication between a researcher and a real person. A real person has unique experiences, opinions, incentives, and emotional reactions. An AI model can approximate those outputs, but it cannot reproduce the actual human experience. The moment you substitute AI survey takers for real humans on a strategic question where the final number is consequential, you have stopped doing research and started running a simulation.

Using AI to pre-screen concepts, prioritize large lists, or provide additional depth of understanding for hard-to-reach quota populations (always calibrated against real human data) is a valid, powerful extension of research. Technology expands the range of what we can do. It must never replace the human voice.

Principle 02.

Validate before you trust.

Every synthetic deployment we recommend requires a held-out human benchmark. This requirement is not optional or negotiable.

The reason is simple: sounding realistic is not the same as being correct. Fluent, well-written AI responses create an illusion of validity. Because the text reads so naturally, researchers can easily fall into a trap of false confidence, even when the underlying data is distorted. Our review of the available evidence is pretty clear: demonstrations that a synthetic model is consistent with human responses on one study does not predict performance on the next.

Requiring a human benchmark does not mean running a full parallel study every single time. It means building a validation component into your study design. This allows the research team to identify where and how the synthetic responses diverge before those results are used to inform a strategic decision.

Principle

03. Rank, don't level.

Use synthetic data for relative ordering and screening. Use real human respondents when you need absolute, stand-alone scores.

This principle is supported by clear, empirical evidence. The landmark Colgate-Palmolive study found that synthetic respondents achieved 90% of human correlation for product rankings, which is a relative ordering task. Similarly, our own restaurant brand tracker found that the relative order of brands was mostly preserved, even while the absolute scores were heavily inflated by 4 to 19 points.

Synthetic survey takers show potential to tell you which flavor idea ranks first, which marketing message is the weakest, or which product concept should be dropped. They cannot reliably predict that sixty-two percent of your target market will recommend your brand. Those stand-alone metrics will be distorted by training data bias in ways that vary by brand and category.

You cannot trust AI survey takers to provide accurate brand awareness, favorability, or purchase intent scores. Reporting these simulated scores as business benchmarks means drawing strategic conclusions that the methodology simply does not support.

Principle

04. Pick the right elicitation method.

How you ask the LLM questions matters more than which LLM you use.

Using direct numeric prompts (such as asking the AI to rate an item on a scale of one to five), the model produced narrow distributions clustered around the middle. This is a common occurrence that strips away the valuable nuances of human opinion.

To unlock 90% accuracy, researchers had to use a specific, advanced approach: they asked the AI to write a free-text response first, and mapped that text back to a rating scale using mathematical similarity.

And there was another vital requirement: demographic persona conditioning. When researchers did not condition the AI personas on age and income, the overall shape of the distribution curves still improved, but the ranking accuracy collapsed to just 50%. In other words, without proper demographic instructions, the AI's ability to put product preferences in the right order was no better than a coin flip.

This has a direct impact on how you evaluate research vendors. The question is not simply "does your model work?" The question is: "What method does your model use to extract ratings, and do you have evidence that it outperforms direct numeric prompting?"

Principle 05.

Choose robustness over novelty.

Workflows validated by peer-reviewed science beat flashy, proprietary techniques.

The market research industry is currently experiencing a wave of AI experimentation that is running well ahead of validation. New models and vendor claims arrive every quarter. At The Directions Group, we are deliberate about where we place our bets.

Methodological best practices beat shortcuts. Proven business impact (the ability to help a team make a confident, validated choice) beats methodological novelty designed to impress a conference audience.

When our pilot data contradicts a vendor claim, we publish it. When the literature identifies a failure mode, we build it into our evaluation process. And if a technique we have invested in turns out to underperform, we say so.

Our evolving roadmap.

We continue to investigate synthetic respondents. As the technology continues to evolve, we are investing in several areas to build a sharper, faster, more validated research pipeline:

In the near-term

We are adding several things to our standard toolkit: Human calibration will be a standard step in all our AI designs, text-to-score mapping will be added to our tools, and we will use our proprietary River panel and past survey waves to retrospectively test and grade the AI's historical accuracy.

In the mid-term

We are building a detailed domain coverage map to identify precisely where AI models have enough training data to be useful, and where data is too thin to trust.

In the long-term

We are designing pre-registered prediction studies on real product launches to hold AI models accountable to real-world business outcomes. That means we are asking the AI to predict outcomes, recording those predictions, and eventually grading them against the results post-launch.

If you'd like to discuss how these findings can apply to your research

Contact us today
www.directionsgroup.com/getting-started

THE DIRECTIONS group