METHODOLOGY PAPER

DATA STRATEGY · SYNTHETIC DATA & REAL RESPONDENTS

Real, Synthetic, *or Both*

A Methodology for Sourcing Decision-Grade Data in the Age of AI

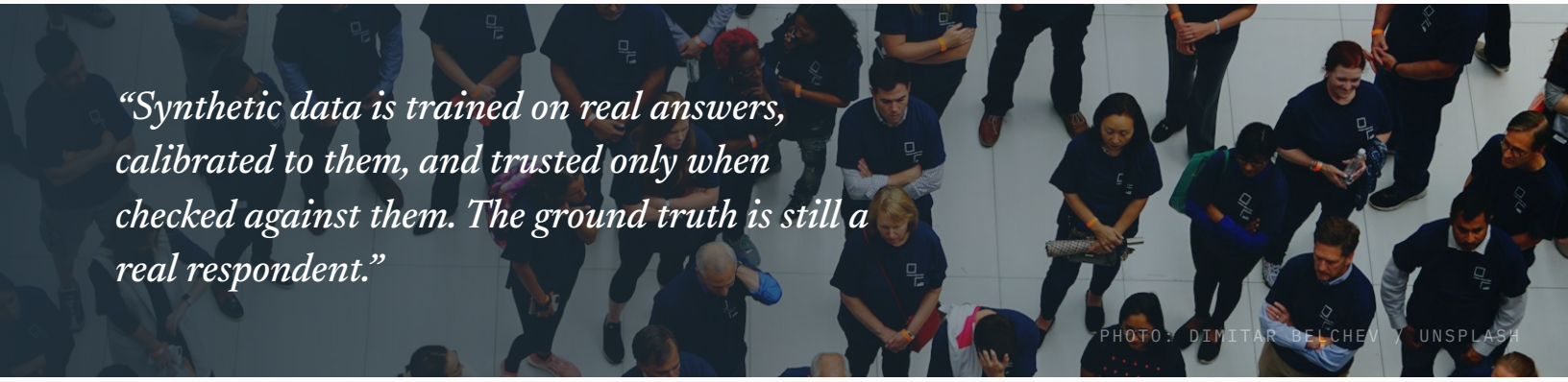
Synthetic data has gone from curiosity to category — augmented sample, LLM "respondents," digital twins, simulated audiences. Used well, it is a genuine accelerator. But it is built from real human answers, calibrated to them, and trustworthy only when validated against them — which is why targeted, quality global panel and CATI interviewing remain the decision-grade ground truth. This is a practical guide to using synthetic data for what it is good at, without mistaking it for the market.

PUBLISHED BY
CatalystMR Research Team

SERIES
Methodology Papers

READING TIME
~18 minutes

EDITION
2026



CATALYSTMR
transforming market research data into marketing solutions

ABSTRACT

Artificial intelligence has made it cheap to generate data that looks like survey responses, and an industry is forming around it: statistical sample **augmentation**, LLM-generated **synthetic respondents**, digital twins, and simulated audiences. For research buyers the noise is loud and the framing is often wrong — "real versus fake." Synthetic data is neither a fraud to be dismissed nor a replacement to be trusted on faith. It is a powerful, fast, increasingly capable **derivative** of real human data.

This paper is a practical, vendor-neutral guide to using it well. It offers a **field guide** to the four kinds of synthetic data; maps **where each genuinely earns its place**; shows **where it breaks**, and why real, verified human data remains the **decision-grade ground truth** that synthetic is trained on, validated against, and bounded by; sets out a framework for choosing to **ask, simulate, or blend**, with governance; and closes on the **complementarity** of the two. The throughline: targeted, quality global panel and CATI are not the alternative to synthetic data — they are the foundation that makes it trustworthy.

CONTENTS

01	The question isn't real vs fake	It's what each kind of data is for
02	A field guide to synthetic data	Four kinds, one common root
03	Where synthetic data earns its place	The real opportunities
04	Where it breaks — and why real is primary	The ground-truth dependency
05	Ask, simulate, or blend	A decision framework & governance
06	The complementarity loop	Real and synthetic, together

WHO THIS IS FOR

Insights and analytics leaders weighing synthetic-data offers; **researchers** designing studies and choosing sample; **data-science and procurement leaders** evaluating synthetic-sample vendors; and **academics** using or assessing LLM-simulated respondents. No technical background is assumed; practitioners can skip to the framework in Section 05.

01 THE PREMISE

The useful question is not "real or fake" — it is "what is each for?"

Synthetic data is no longer hypothetical. Industry analysts forecast it will supply most AI training data within a few years; the largest panel and platform companies are shipping synthetic respondents; and the category now spans everything from statistical sample-boosting to LLM-generated personas. The unhelpful reaction is to argue authenticity. The practical question is narrower: which questions can a model trained on yesterday's human data answer well — and which still demand a real, verified human today?

By 2030

Gartner forecasts synthetic data will overtake real data in AI-model training

GARTNER · FORECAST

By 2027

Gartner forecasts **60%** of data & analytics leaders will hit critical failures managing synthetic data

GARTNER · FORECAST

Model collapse

Training AI recursively on synthetic data degrades it — the real distribution's tails disappear

NATURE · SHUMAILOV ET AL. 2024[3]

Two truths sit behind those figures. Synthetic data is **real enough to be useful** — and **derivative by construction**. Every synthetic respondent is produced by a model that learned from real human responses; it interpolates what it has seen, it does not observe the world anew. That is why the same momentum that makes synthetic data attractive also makes **real, verified human data more valuable, not less**: it is the ground truth that trains the models, the benchmark that validates them, and the renewable source that keeps them from drifting away from reality.

> THE REFRAMING
>
> The choice is not real or synthetic. It is knowing which questions a model trained on past human data can answer well, and which require asking a real person now. **Real, verified data is the ground truth** — it trains synthetic data, validates it, and sets its ceiling.

02 THE LANDSCAPE

Four kinds of synthetic data — one common root.

"Synthetic data" is one phrase for several very different things, and conflating them is how buyers get burned. The four below differ in how they are made and what they can be trusted for — but all share one root: each is generated by a model that learned from real data. Knowing which kind a vendor is selling is the first act of due diligence.

KIND	WHAT IT IS	BUILT FROM	BEST USE
AUGMENTATION **Statistical "boosting"**	Models trained on a study's **own real data** generate extra respondents for thin or niche cells. Quantitative, close-ended, checkable.	Your real survey data, **per study**	Stabilising **small subgroups**; verifiable against a real holdout
GENERATION **LLM respondents & digital twins**	Language models role-play personas to answer surveys, rank concepts, or write open-ends from scratch.	Pre-trained model weights + persona prompts	**Exploration**, pre-testing, structured ranking — not statistical inference
PRIVACY **Privacy-preserving synthetic**	Artificial records that keep a dataset's statistical patterns while containing **no real PII**.	A real dataset + a generative model	Sharing, training & testing **without exposing** personal data
SIMULATION **Synthetic controls & scenarios**	Model-generated data for conditions or rare events not present in the real sample.	Real data + rules / simulation	**What-if** analysis, edge cases, stress-testing models

WHY THE DISTINCTION MATTERS

Statistical **augmentation** — trained on a study's own data and checkable against a holdout — is a very different proposition from an **LLM persona** answering from pre-trained weights. The first extends real data; the second improvises in its absence. A buyer who hears only "synthetic respondents" cannot tell which they are getting — so the first question is always: **made how, from what data, and validated against what?**

03 THE OPPORTUNITIES

Used for what it is good at, synthetic data is a real accelerator.

A method partner that only argued against synthetic data would not be worth reading. Used well, it removes genuine friction — and the evidence that it can work, under the right conditions, is real. The point is to match the tool to the task.

Where it genuinely helps

- **Augmenting thin cells** — stabilising small, hard-to-reach, or niche subgroups by boosting a real subsample.
- **Pre-testing & concept screening** — pressure-testing questionnaires and ideas before going to field.
- **Hypothesis generation** — fast, cheap directional reads to decide what is worth asking real people.
- **Simulation** — modelling rare events and what-if scenarios the real sample can't supply.
- **Privacy** — sharing and testing on data that holds the patterns but none of the PII.

What it's genuinely good at

Synthetic respondents perform best on **structured tasks** — ranking, pricing, sentiment — and, as augmentation, on stabilising subgroups when trained on solid real data. Peer-reviewed work shows language models conditioned on real human backgrounds can emulate subgroup response patterns with surprising fidelity,[1] and data-science research has found **no significant difference** between some predictive models built on well-made synthetic data and on real data.[4] The opportunity is real; so are its limits.

> BUYER'S QUESTION
>
> Ask: **"Is this directional, exploratory, or low-stakes — or decision-grade?"** Synthetic shines on the first; the second still needs real respondents.

04 THE DEPENDENCY

Synthetic data inherits the limits of what it learned from.

Every kind of synthetic data sits on a foundation of real human responses. That dependency is also its constraint: a model can only reflect the world it was trained on, and pushed past that, it averages, distorts, and drifts. Read the stack from the bottom up.

LAYER 3 · OUTPUT

Synthetic respondents & augmented sample

Only as representative as the data beneath them — and prone to **under-dispersion, bias amplification, and drift** when pushed past it.

▲ TRAINED ON · VALIDATED AGAINST

LAYER 2 · MODEL

Models trained & calibrated

Learn patterns from real responses; they **interpolate what they have seen** — they do not observe the world anew.

▲ LEARNED FROM

LAYER 1 · GROUND TRUTH

Real, verified human data – panel & CATI

The only layer that observes reality directly. Remove it and the stack has nothing to stand on — trained recursively on its own output, a model **collapses**.

WHAT THE RESEARCH SHOWS

Synthetic survey data can match real averages while showing **less variation, distorted relationships between variables, sensitivity to small prompt changes, and drift over time**[2] — and a landmark Nature study shows that models trained recursively on synthetic data suffer **"model collapse,"** losing the tails of the real distribution.[3] Both point one way: real human data is the irreplaceable anchor, and synthetic output must be **validated against a real holdout**, never trusted on its face.

05 THE FRAMEWORK

Match the method to the decision — and govern it honestly.

The practical question is rarely all-or-nothing. It is which method fits this decision, given its stakes and the data you already hold — and how to use synthetic data without quietly lowering the evidentiary bar.

Match method to the decision

- **Ask** (panel / CATI) when the call carries real risk, the audience is niche or must be verified, or you need novelty, nuance, or ground truth to train and validate against.
- **Simulate** (synthetic) for exploration, pre-testing, what-ifs, and boosting thin cells where a directional answer is enough.
- **Blend** — augment a real subsample, or pre-test synthetically then confirm with real respondents — as the common, defensible middle.

Govern it honestly

Disclose where synthetic data is used and how it was made; **validate** it against a real holdout; and demand **independent** validation rather than vendor-selected metrics. Conduct and quality frameworks apply to synthetic data as to any other method.[5,6] The governance risk is real: analysts forecast that most data leaders will struggle to manage synthetic data well.

> BUYER'S QUESTION
>
> Ask: **"How was this validated against real respondents — and by whom?"** If the only proof is the vendor's own, treat the result as a hypothesis, not a finding.

06 TOGETHER

Real and synthetic are a loop — with real data at its centre.

The strongest programs do not choose once and for all; they run a cycle in which real and synthetic data feed each other — and the cycle only holds because real, verified responses anchor it at both ends.

START · GROUND TRUTH
Real human data
Targeted panel & CATI capture verified, decision-grade responses.

→

TRAIN
Models learn
Augmentation and persona models are trained and calibrated on that real data.

→

EXTEND
Synthetic scales it
Boost thin cells, pre-test concepts, simulate what-ifs — fast and cheap.

→

CONFIRM
Decisions of consequence
Returned to real respondents for the calls that carry real risk.

↻ EVERY SYNTHETIC STEP IS **VALIDATED AGAINST** — AND REFRESHED BY — REAL HUMAN DATA: THE GROUND TRUTH THAT CLOSES THE LOOP

Real and synthetic are not rivals; they are a loop with real data at its centre. **The better and more representative the real sample, the better the synthetic that learns from it** — which is why investing in targeted, quality global panel and CATI is not the alternative to synthetic data, but the foundation that makes it trustworthy. The future is hybrid: augmented designs, continuous validation, and synthetic exploration — all anchored to a renewable supply of real human truth.

WHERE THIS SERIES GOES DEEPER

No. 142 · 137	**Validation & QC.** Confirming a real respondent is who they claim, and the eleven-point quality framework that keeps the ground truth clean.
No. 138	**Fraud & AI detection.** Keeping AI-generated and fraudulent answers out of the real sample that trains everything downstream.
No. 147 · 132	**Verified credentials & physician sample.** Why targeted, verified human sample is decision-grade in the first place.

CATALYSTMR
transforming market research data into marketing solutions

▲ CONCLUSION

Real data is the ground truth. Synthetic data is its multiplier.

Synthetic data has earned a permanent place in the research toolkit — for augmenting thin cells, pre-testing, simulation, and privacy. Used for those jobs, and validated honestly, it is fast, scalable, and genuinely useful. But it is a derivative: trained on real human responses, calibrated to them, and bounded by them — and recursive over-reliance degrades it. So the rise of synthetic data does not diminish the real respondent; it raises the value of a clean, representative, decision-grade ground truth. Choose to ask, simulate, or blend by the stakes of the decision; disclose and independently validate what is synthetic; and treat targeted, quality global panel and CATI not as the alternative to AI, but as the foundation it stands on. The ICC/ESOMAR Code and the ISO 20252 framework let buyers ask for that rigour in consistent terms.[5,6]

§ REFERENCES

[1] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. **"Out of One, Many: Using Language Models to Simulate Human Samples."** Political Analysis 31, no. 3 (2023): 337–351.

[2] Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. **"Synthetic Replacements for Human Survey Data? The Perils of Large Language Models."** Political Analysis 32, no. 4 (2024): 401–416.

[3] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. **"AI Models Collapse When Trained on Recursively Generated Data."** Nature 631 (2024): 755–759.

[4] Patki, N., Wedge, R., & Veeramachaneni, K. **"The Synthetic Data Vault."** IEEE Int'l Conf. on Data Science and Advanced Analytics (DSAA), 2016.

[5] ICC/ESOMAR. **International Code on Market, Opinion and Social Research and Data Analytics.** ICC & ESOMAR.

[6] International Organization for Standardization. **ISO 20252:2019 — Market, opinion and social research, including insights and data analytics.** Geneva: ISO, 2019.

Sourced figures. The "by 2030 / by 2027 / 60%" figures are **third-party Gartner forecasts** reported in the trade and technology press; they are forward-looking projections, not measurements or CatalystMR data, and are cited as industry context. The "model collapse" finding is from the peer-reviewed Nature study [3]. No CatalystMR or study-specific rate is published in this paper.

§ ABOUT

CatalystMR

CatalystMR is a global market-research panel and fieldwork partner specialising in **targeted, verified sample** across healthcare, B2B, and niche audiences, delivered by **online panel and live CATI interviewing**. We treat real, decision-grade human data as the ground truth — and advise honestly on where synthetic augmentation or simulation adds value around it.

Compliance posture: aligned to the ESOMAR Code and Guidelines and the ISO 20252 framework; certified under the EU-U.S., UK, and Swiss Data Privacy Frameworks, with personal data siloed from response data.

SYNTHETIC DATA · PANEL & CATI · DATA QUALITY · AI IN RESEARCH · ESOMAR · ISO 20252

This paper cites third-party academic and industry sources for the capabilities and limits of synthetic data; named vendors and platforms are industry examples, not endorsements. Synthetic-data techniques and the market around them are evolving quickly; claims here are current to the 2026 edition and should be re-checked over time. "Aligned to ISO 20252" denotes framework conformance, not third-party certification; any turnaround estimate is a modelled feasibility range, not a guarantee.

Tell us the decision you're making and the data you hold — we'll help you choose where real panel or CATI is essential, where synthetic augmentation or simulation adds value, and how to validate the blend — with a modelled feasibility range, typically within 24 hours.

sales@catalystmr.com
1-800-819-3130
catalystmr.com